INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO
UNIVERSIDADE TÉCNICA DE LISBOA

EXPLORANDO DADOS MULTIVARIADOS
ANÁLISE GRÁFICA - TÓPICOS E APLICAÇÕES

Isabel Maria Bento
de Brito Bouzerb

Lisboa, Junho de 1994

INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO

UNIVERSIDADE TÉCNICA DE LISBOA

EXPLORANDO DADOS MULTIVARIADOS: ANÁLISE GRÁFICA - TÓPICOS E APLICAÇÕES

Tese de mestrado em Matemática Aplicada à Economia e Gestão realizada sob orientação do Professor Doutor Bento Murteira

Isabel Maria Bento

de Brito Bouzerb

Lisboa, Junho de 1994

Gostaria de deixar expresso o meu profundo agradecimento ao Professor Doutor Bento Murteira que prestou uma dedicada e preciosa orientação a este trabalho e auxiliou na recolha de bibliografia e execução de alguns gráficos.

Queria, ainda, agradecer aos meus colegas que, de alguma maneira, contribuiram com informações pertinentes melhorando a realização do trabalho.

O meu reconhecimento vai, também, para todos os que, à minha volta, demonstraram a sua compreensão e apoio e sem os quais o trabalho não teria sido possível.

Por último, devo esclarecer que os erros e incorrecções são da minha exclusiva responsabilidade.

QUADRO DE MATÉRIAS

1. INTRODUÇÃO 4

2. CONCEITOS FUNDAMENTAIS 11

3. ESQUEMAS DE REPRESENTAÇÃO DIRECTA E UNILATERAL 22

3.1. REPRESENTAÇÃO SIMBÓLICA 24

3.2. REPRESENTAÇÃO FUNCIONAL: AS CURVAS DE ANDREWS 54

4. REPRESENTAÇÃO DE MATRIZES DE PROXIMIDADE 71

4.1. "*MULTIDIMENSIONAL SCALING*" CLÁSSICO 75

4.2. "*MULTIDIMENSIONAL SCALING*" ORDINAL 95

5. ESQUEMA DE REPRESENTAÇÃO BILATERAL:

"*BIPLOT*" 115

6. ESQUEMA DE REPRESENTAÇÃO TRILATERAL:

"*MULTIDIMENSIONAL SCALING*" DE DIFERENÇAS INDIVIDUAIS .. 142

7. COMPARAÇÃO DE REPRESENTAÇÕES:

ANÁLISE PROCRUSTEANA 161

8. CONCLUSÃO 178

ANEXO 180

BIBLIOGRAFIA 195

" Uma imagem vale mil palavras "

"Qual é a utilidade de um livro,

pensou a Alice, se não tiver gravuras?"

- Lewis Carrol - "*Alice no País das Maravilhas*".

CAPÍTULO 1

INTRODUÇÃO

A actualidade é uma época de obsessão pelos números. Qualquer domínio da actividade humana usa e abusa da informação numérica e estatística; mesmo em disciplinas como História ou Linguística os métodos quantitativos são, hoje, exigidos. A esta mentalidade alia-se o enorme aumento da capacidade tecnológica para processar todo o manancial de informação, permitindo aguçar, em *crescendo*, o apetite pelas quantificações.

Krzanowski(1990) faz uma listagem de exemplos encontrados num conjunto de áreas diversificadas onde é importante a informação estatística.

> *"Na Agricultura podemos querer classificar os tipos de solos, desenvolver novas variedades de sementes ou determinar as combinações de características das sementes que melhor respondem ao tratamento com diferentes fertilizantes. Em Antropologia pode ser interessante distinguir entre diferentes populações Homínidas ou*

identificar a que populações pertencem alguns restos de esqueletos de crânios. Em Educação podemos querer explicar a diferença entre as classificações académicas dos estudantes ou relacionar a sua situação familiar com o desempenho escolar. Na Indústria, o esforço pode ser canalizado para relacionar custos de produção e custos de material ou relacionar o sucesso da empresa com o ambiente económico ou, ainda, determinar os factores que conduzem ao melhor processo de produção. Em Marketing é habitual querer classificar consumidores, relacionar novos produtos e produtos já estabelecidos ou explicar as diferenças entre vários produtos no mercado. Em Medicina tem aumentado a procura de um suporte estatístico sobretudo para problemas de diagnóstico. O importante é distinguir pacientes doentes e saudáveis ou identificar os sintomas dos pacientes. Psicologia é uma área na qual a análise multivariada tem uma longa tradição e onde os problemas típicos são a medição da consistência de afirmações sensoriais de juízes ou a determinação de factores que afectam a habilidade cognitiva das crianças. Em Sociologia podemos querer classificar respostas a um questionário, representar semelhanças entre tribos linguísticas ou determinar os factores que afectam o comportamento dos delinquentes. Finalmente, Taxonomia é uma área que cobre a maioria das Ciências Biológicas e tem interesse em distinguir populações de organismos

diferentes ou classificar espécies de flora."

O produto final da recolha de informação quantitativa é um conjunto multivariado de dados no qual para um número *n* de indivíduos ou casos são medidas *p* variáveis. O investigador tem de dar sentido a esta massa de números esperando encontrar algumas características ou padrões de comportamento interessantes.

Os métodos usados na procura sistemática do conhecimento percorrem as mesmas etapas ou etapas muito semelhantes em todos os ramos da ciência. Envolvem o reconhecimento e formulação de problemas, a recolha de dados empíricos relevantes, através da observação passiva ou através da experiência e o uso de análise estatística para explorar as relações entre os dados ou para testar hipóteses específicas àcerca das observações.

A mensagem estatística pode ser divulgada através de palavras, tabelas ou gráficos. A eficiência da transmissão de cada uma das vias depende do conteúdo da mensagem e elas devem complementar-se. Se as palavras traduzem, por vezes, a realidade de uma forma complexa, as tabelas podem fazê-lo de forma concisa. Os gráficos, por seu turno, são essenciais para transmitir relações e tendências das leis estatísticas de um modo visual simplificado e informal.

Refere Gordon(1980):

> *" O Homem é, por excelência, um animal geométrico que necessita e deseja figuras para simplificar e ao mesmo tempo estimular o conhecimento profundo. "*

O engenho de *olhar* para os dados é frequentemente um grande passo na resposta a muitas questões; por isso, os métodos de representação de conjuntos de dados formam um primeiro estádio vital em qualquer análise sistemática. Tukey, citado por Kulkarni e Paranjage(1984), faz a seguinte comparação:

> *"A relação entre os gráficos e as técnicas formais de inferência é análoga à relação entre o polícia-detective e o juíz. Os procedimentos gráficos são a parte correspondente do trabalho do detective na procura de pistas para desvendar os mistérios dos dados. Os instrumentos numéricos de inferência da estatística clássica são a contrapartida do juíz que, no tribunal, pondera o grau de evidência das pistas para determinar a confiança que lhes deve atribuir."*

Está amplamente reconhecido que a análise de dados se reparte em duas etapas. A primeira envolve a exploração dos dados numa tentativa de reconhecer qualquer estrutura não aleatória e gerar hipóteses interessantes para estudo futuro. Nesta etapa encontrar

a pergunta é frequentemente mais importante do que encontrar a resposta, pelo que, não se tornam necessários modelos formais para determinar respostas específicas a questões rígidas. Em vez deles desejam-se técnicas que antecipem possíveis estruturas transmitidas pelos dados abrindo caminho a um largo conjunto de explicações alternativas. Estas técnicas caracterizam-se pela importância que atribuem às apresentações gráficas e visuais e pela falta de um modelo estocástico associado, de modo que, questões de significância estatística, dificilmente se colocam. A análise confirmatória constitui a segunda etapa e principia no momento em que o investigador tem uma hipótese bem definida, constrói modelos e efectua testes que lhe permitam aceitar ou rejeitar a hipótese.

A análise multivariada exploratória é, assim, rica em métodos gráficos informais destinados a pesquisar espaços multidimensionais e a exibir a estrutura dos dados a partir da qual se formularão hipóteses e modelos para investigação futura. As palavras-chave da análise exploratória são *simplificação* e *apresentação*. Por outras palavras, pretende sumariar-se um conjunto grande de dados através de, relativamente, poucas medidas. Em muitas análises uma descrição meramente informativa é, mesmo, suficiente.

Um grande vulto da estatística, Fisher, citado por Everitt(1978), sintetiza as ideias expostas:

"A análise preliminar da maioria dos dados é facilitada pelo uso de diagramas. Os diagramas não provam nada mas tornam as características dos dados facilmente apercebíveis; não são substitutos para os testes de significância mas têm um valor precioso na sugestão desses testes e na explicação das conclusões."

O presente trabalho tem como objectivo examinar métodos possíveis de representação gráfica de dados multivariados no âmbito da análise exploratória. A abordagem aqui traçada é orientada no sentido do problema, querendo com isto significar que as técnicas expostas aparecem como resposta ao tipo de problema concreto colocado nas mãos dos investigadores, sobretudo o tipo de dados recolhidos, e não agrupadas segundo o esquema teórico que perfilham. No capítulo 2 tecem-se considerações fundamentais para a compreensão do que se escreve em seguida. No capítulo 3 descrevem-se métodos muito simples de representação das variáveis ou dos indivíduos; no capítulo 5 efectuam-se configurações simultâneas dos indivíduos e das variáveis e no capítulo 6 desenham-se configurações que comportam os indivíduos e as variáveis para mais que uma amostra da mesma população. O capítulo 4 dedica-se à representação de relações que se estabelecem entre indivíduos e, finalmente, no capítulo 7 comparam-se configurações obtidas através das técnicas anteriores.

Não se pretende, neste espaço descrever até ao esgotamento todas as técnicas existentes de representações gráficas mas tão só propôr e aplicar, principalmente, alguns dos métodos geométricos de classificação.

É impensável proceder a análise de dados multivariados sem usar aplicações informáticas. Foi utilizado na elaboração deste trabalho o seguinte software: AXUM, GENSTAT, SPSS e STATGRAPH.

CAPÍTULO 2

CONCEITOS FUNDAMENTAIS

Vários termos são usados na literatura para as entidades do conjunto de dados. As *unidades* de experimentação ou de observação denominam-se, genericamente, *indivíduos* ou, mais especificamente, sobretudo em Psicologia, Sociologia ou contextos experimentais, *sujeitos*. As medições feitas em cada unidade são conhecidas por *variáveis*, mas também por *respostas*, *atributos* ou *estímulos*.

O conjunto de dados multivariados é habitualmente exibido como uma matriz cujas linhas se referem aos indivíduos e as colunas às variáveis medidas para cada indivíduo. Simbolicamente:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

em que x_{ij} é o valor da j-ésima variável medida para o i-ésimo indivíduo. A $\mathbf{X}$ chama-se, daqui em diante, matriz de dados ou

matriz de observações.

Os vários procedimentos actualmente disponíveis para representar graficamente dados multivariados dividem-se em duas categorias, a saber: abordagem directa e abordagem de projecção e optimização.

A primeira, é uma simples generalização das familiares técnicas de representação univariadas tais como diagramas de barras, polígonos de frequência, etc. Basta escolher uma curva ou símbolo apropriado para representar cada indivíduo ou variável e de seguida desenhá-los em duas dimensões. Os eixos utilizados nestas representações não têm, habitualmente, relevância directa no problema em estudo.

A segunda, envolve um raciocínio geométrico e subdivide-se em duas vertentes. Por uma parte, assume, tacitamente, que existe uma representação *verdadeira* dos dados muito complicada, isto é, num grande número de dimensões, pelo que não pode ser observada directamente. Em substituição encontra-se uma aproximação projectando a verdadeira configuração num pequeno número de dimensões correctamente escolhidas. Neste caso a interpretação dos eixos tem grande importância no problema em estudo. Por outra parte, assume que não existe uma única representação verdadeira mas que se pretende construir a representação que melhor se aproxime dos objectivos específicos. Em essência, a abordagem de projecção e optimização visa reduzir o número de dimensões,

preservando, claro, as características dos dados.

Para que a configuração aproximada transmita com confiança a estrutura das entidades em estudo é necessário que entidades semelhantes sejam representadas por pontos situados próximos uns dos outros e entidades dissemelhantes por pontos distantes. A ideia básica é a da relação que envolve dissemelhança de entidades e distância entre pontos de uma configuração.

O conceito de dissemelhança conduz a um segundo tipo de matrizes em que as linhas e as colunas se referem às mesmas entidades e os elementos representam relações entre pares de entidades. Assim, a matriz

$$\Delta_I = \begin{bmatrix} \delta_{11} & \delta_{12} & \dots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \dots & \delta_{2n} \\ \dots & \dots & \dots & \dots \\ \delta_{n1} & \delta_{n2} & \dots & \delta_{nn} \end{bmatrix},$$

em que o elemento genérico δ_{ij} $(i,j=1,\dots,n)$ é o valor da dissemelhança entre o i-ésimo e o j-ésimo indivíduos, contém informação de *proximidades* entre indivíduos. A matriz

$$\Delta_V = \begin{bmatrix} \delta^*_{11} & \delta^*_{12} & \dots & \delta^*_{1p} \\ \delta^*_{21} & \delta^*_{22} & \dots & \delta^*_{2p} \\ \dots & \dots & \dots & \dots \\ \delta^*_{p1} & \delta^*_{p2} & \dots & \delta^*_{pp} \end{bmatrix},$$

em que δ^*_{kl} $(k,l=1,\dots,p)$ é o valor da dissemelhança entre a k-ésima e a l-ésima variáveis, contém informação de ***associação*** entre variáveis.

Ao longo dos anos têm sido propostas medidas de dissemelhança para, a partir da matriz **X**, construir Δ_I. As medidas de dissemelhança mais comuns são as distâncias ou métricas (é usual designar distância por d, pelo que se fará $\delta=d$) que se caracterizam pelas seguintes propriedades:

i) $d_{ij} \begin{cases} = 0 \text{ se } i=j \\ > 0 \text{ se } i\neq j \end{cases}$,

ii) $d_{ij} = d_{jl}$ (simetria),

iii) $d_{ij} \leq d_{ik} + d_{kj}$ (desigualdade triangular).

Alguns investigadores, na área de psicologia, tendem a usar a palavra 'métrica' como referência à escala numérica que a dissemelhança está medida e não como forma de dizer que satisfaz as condições acima.

Existem diversas medidas de distância, sendo as mais conhecidas:

1) Métrica de Minkowski,

$$d_{ij} = \left[\sum_{l=1}^{p} |(x_{il}-x_{jl})|^{\lambda} \right]^{1/\lambda} \quad \text{com } \lambda \text{ inteiro.}$$

Quando $\lambda=1$ ou $\lambda=2$ tem a designação particular de, respectivamente, distância "*City Block*" e distância euclideana.

2) Distância Euclideana ponderada,

$$d_{ij} = \left[\sum_{l=1}^{p} w_l (x_{il}-x_{jl})^2 \right]^{1/2} .$$

Quando w_l é o inverso da variância da l-ésima variável tem a designação particular de distância de Pearson ou do χ^2.

3) Distância de Mahalanobis,

$$d_{ij} = [(\mathbf{x}_i-\mathbf{x}_j)\mathbf{S}^{-1}(\mathbf{x}_i-\mathbf{x}_j)^T]^{1/2} ,$$

com $\mathbf{x}_i=(x_{i1},x_{i2},\ldots)$, $\mathbf{x}_j=(x_{j1},x_{j2},\ldots)$ e $\mathbf{S}$ a matriz de variâncias-covariâncias de $\mathbf{X}$.

Por vezes é calculada a semelhança podendo a dissemelhança ser obtida através de uma transformação monótona decrescente. As medidas de semelhança (designa-se semelhança por c) caracterizam-se por:

i) c_{ij} $\begin{cases} = 1 \text{ se } i=j \\ > 0 \text{ e } < 1 \text{ se } i \neq j \end{cases}$,

ii) $c_{ij} = c_{ji}$ (simetria).

As medidas de semelhança e de dissemelhança estão firmemente relacionadas, mas em sentido contrário, sendo possível transformar uma distância numa semelhança através de, por exemplo,

$$c_{ij} = (1+d_{ij})^{-1}. \tag{2.1}$$

O processo inverso não é tão óbvio devido à desigualdade triangular que qualquer distância deve satisfazer mas pode basear-se na transformação seguinte,

$$d_{ij} = 2(1-c_{ij})^{1/2} . \tag{2.2}$$

Existem muitas outras medidas que seria fastidioso enumerar (consulte-se Coxon(1982) a este propósito). A escolha da medida a usar depende do tipo de dados recolhidos, de alguma reflexão preliminar e da sensibilidade para o fenómeno em causa.

As correlações entre variáveis são uma espécie de semelhança, existindo várias transformações apropriadas para converter correlações ou covariâncias em dissemelhanças.

Em muitos casos não se conhece a matriz **X** e observam-se directamente as dissemelhanças como, por exemplo, em estudos psicológicos onde os dados constituem uma matriz de opiniões àcerca de pares de estímulos.

A análise exploratória pode ser dirigida às colunas da matriz **X** visando salientar as características das variáveis ou ser dirigida às linhas com o propósito de descobrir relações entre indivíduos. Por vezes a distinção entre os dois tipos de técnicas não é muito clara. A dicotomia variáveis/indivíduos é, ainda, visível se a análise recair sobre Δ_V ou Δ_I e dá origem a dois espaços: o espaço das variáveis e o espaço dos indivíduos. As técnicas multivariadas que têm como objecto Δ_I designam-se *técnicas-Q* e as que se baseiam em Δ_V designam-se *técnicas-R*.

Qualquer representação gráfica está restringida ao facto de ser elaborada numa folha de papel e, assim, por aspectos práticos

terá de ser bidimensional. Há, claro, excepções. É perfeitamente possível estender os conceitos geométricos do plano a construções tridimensionais físicas ou desenhadas em perspectiva. Porém a representação a duas dimensões, única usada neste trabalho, é, ainda, dominante.

Na prática existem duas grandes categorias de dados: *quantitativos* e *qualitativos*. Os primeiros atribuem um valor numérico a cada indivíduo observado e subdividem-se em *discretos* ou *contínuos*. Os segundos fazem uma classificação dos objectos em estratos que descrevem as qualidades possuídas por cada objecto e são medidos, quer numa escala *nominal*, quer numa escala *ordinal*.

No âmbito da abordagem de projecção e optimização dá-se ênfase a alguns métodos de classificação cuja solução resulta numa representação gráfica, denominados, por esse motivo, modelos geométricos. O "*Scaling*" clássico e o "*Scaling*" ordinal são dois métodos para obter uma configuração geométrica de pontos em que as distâncias entre os pontos se aproximam, de modo específico, das dissemelhanças entre os correspondentes pares de objectos. Os dois métodos diferem em muitos aspectos, particularmente no modelo que relaciona distâncias e dissemelhanças. O "*Scaling*" de diferenças individuais é um método mais formal que analisa a situação na qual juízes diferentes avaliam o mesmo conjunto de objectos e constrói uma representação geométrica dos objectos que reflecte as diferenças entre eles tal como os juízes as

entenderam. As técnicas de "*Scaling*" têm sido largamente utilizadas e descritas com diferente terminologia: em Ecologia denominam-se "Métodos de Ordenação", em Psicologia "*Multidimensional Scaling*" e em Arqueologia "Métodos de Seriação". O "*Biplot*" efectua uma só representação geométrica conjunta dos objectos e das variáveis possibilitando investigar as relações entre os dois conjuntos. Finalmente, a Análise Procrusteana compara diferentes configurações geométricas de um mesmo conjunto de objectos, permitindo avaliar a diferença entre as representações.

No âmbito da abordagem directa relatam-se, neste trabalho, técnicas de representação simbólica - perfis, glyphs, caras de Chernoff, etc. - e funcional - curvas de Andrews.

Análises de conglomerados, de componentes principais e de correspondências são temas que proporcionam, igualmente, representações gráficas mas não são tratados no presente estudo.

Um conjunto de observações será utilizado ao longo do texto para aplicação dos métodos expostos. O Exemplo - Quadro 2.1 - foi retirado de Santos(1986). As variáveis têm o seguinte significado:

CPPNB - Capitação do PNB, em dólares dos EUA,

PIBA - Percentagem do PIB originado na agricultura,

QUADRO 2.1 - INDICADORES DO NÍVEL DE DESENVOLVIMENTO ECONÓMICO PARA 30 PAÍSES DA EUROPA E DE ÁFRICA NO ANO DE 1980

OBS	PAÍSES	CPPNB	PIBA	ENRG	PURB	EPVN	HMED
1	Alemanha	13590	2	6053	85	73	450
2	Áustria	10230	4	5102	54	72	400
3	Bélgica	12180	2	7431	72	73	400
4	Burkina Faso	210	40	33	10	39	48510
5	Burundi	200	55	16	2	42	45020
6	Camarões	670	27	154	35	47	13670
7	Dinamarca	12950	4	7971	84	75	480
8	Egipto	580	30	595	45	57	970
9	Espanha	5400	8	2914	74	73	460
10	Etiópia	140	51	25	14	40	58490
11	Finlândia	9720	0	6351	62	73	530
12	França	11730	4	5368	78	74	580
13	Gana	420	60	268	36	49	7630
14	Grécia	4380	16	2605	62	74	420
15	Holanda	11470	4	8068	76	75	540
16	Itália	6480	6	3725	69	73	340
17	Libéria	530	36	502	33	54	9610
18	Malawi	230	58	59	10	44	40950
19	Marrocos	900	23	368	41	56	11200
20	Nigéria	1010	63	169	13	49	12550
21	Noruega	12650	5	11928	53	75	520
22	Portugal	2370	13	1822	31	71	540
23	Quénia	420	38	208	14	55	10500
24	Reino Unido	7920	2	5363	91	73	650
25	Ruanda	200	40	28	4	45	31510
26	Serra Leoa	280	31	166	22	47	18280
27	Sudão	410	38	101	25	46	8800
28	Suécia	13520	3	5223	87	75	490
29	Tunísia	1310	17	652	52	60	3690
30	Zaire	220	30	107	34	47	14780

ENRG - Consumo de energia por habitante em quilogramas de equivalente carvão,

PURB - Percentagem da população urbana,

EPVN - Esperança de vida à nascença,

HMED - Número de habitantes por médico.

Resta escrever, nestas notas preliminares, que os resultados apresentados ao longo do texto e referenciados com *, são demonstrados no capítulo 9.

CAPÍTULO 3

ESQUEMAS DE REPRESENTAÇÃO DIRECTA E UNILATERAL

A análise gráfica exploratória de um conjunto de dados multivariados tem como objectivo identificar, de um modo informal, agrupamentos, padrões de comportamento ou *outliers*. A ideia é ter um primeiro contacto com a estrutura das observações e um simples gráfico torna-se prelúdio essencial para qualquer análise estatística.

Os dados multivariados podem sempre ser apresentados sob a forma de uma matriz. No presente capítulo as linhas e as colunas da matriz definem entidades diferentes, respectivamente, os n sujeitos e as p variáveis.

Com as técnicas de representação directa as observações são apresentadas em todas as suas vertentes, não havendo lugar, portanto, a qualquer redução prévia do número de dimensões. Os métodos directos incluídos neste ponto socorrem-se da simbologia para apresentar os dados utilizando pontos, curvas, estrelas, caras, etc.

Uma vez que os dados se apresentam dispostos matricialmente, os esquemas de representação que a seguir se dão conta, são apropriados para "ler" as linhas da matriz. Estes métodos salientam as características de cada sujeito retratadas nos valores assumidos pelas variáveis. Podem, igualmente, "ler" as colunas da matriz e, neste caso, dão conta das relações entre variáveis. A denominação de técnica *unilateral* advém desta posição de "leitura" dirigida apenas a uma entidade da matriz.

Nas últimas duas ou três décadas um número considerável de técnicas unilaterais, bastante expressivas, tem sido desenvolvido. Algumas destas técnicas, em especial as curvas de Andrews e as faces de Chernoff, entusiasmaram os investigadores, e gozam de grande popularidade devido à sua simplicidade.

Du Toit, Steyn e Stumpf(1986) relatam que vários estudos foram feitos com vista a determinar qual a melhor técnica, de entre as discutidas neste capítulo, para salientar as diferenças entre as observações multivariadas. Concluíram que as curvas de Andrews são o método com mais sucesso, logo seguido dos rostos de Chernoff, enquanto os perfis se revelaram com menos êxito e colocaram no mesmo nível os glyphs, as estrelas e as caixas.

Os métodos directos, se bem que apresentem sérias desvantagens, constituem instrumentos gráficos preciosos para "espreitar" os

conjuntos de dados multivariados.

3.1 REPRESENTAÇÃO SIMBÓLICA

OS *RABISCOS DO DESENHADOR (DRAFTSMAN' DISPLAY)*

Representar dados univariados ou mesmo bivariados é uma tarefa bastante simples e sobejamente conhecida. Se se investiga apenas uma variável, os familiares gráficos de barras, polígonos de frequência ou histogramas constituem algumas representações possíveis. No caso de duas variáveis é comum construir um gráfico de dispersão medindo cada variável em cada um dos eixos. Quando o número de variáveis, seja p, presente nos dados é superior a dois surge a possibilidade de originar um gráfico de dispersão a partir de cada par de variáveis. Esta técnica, que funciona no espaço das variáveis, é a extensão natural a p dimensões da representação bivariada.

A configuração resultante tem um aspecto que em muito se assemelha aos rabiscos de um desenhador. Variações sobre o tema são bem delineadas por Tukey e Tukey(1981) e ainda por Chambers, Cleveland, Kleiner e Tukey(1983).

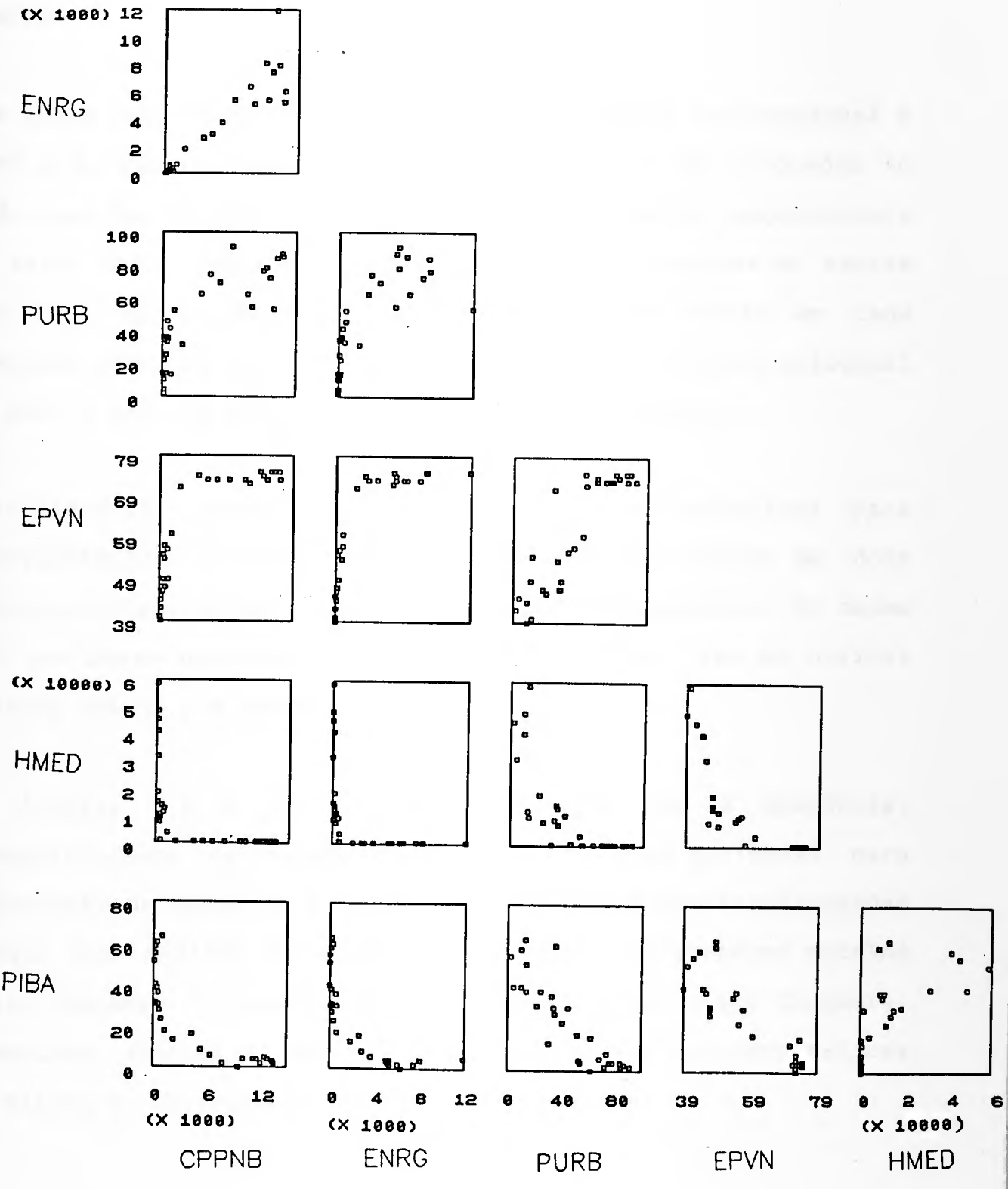
É conveniente arranjar os diagramas de modo que gráficos

adjacentes tenham um eixo em comum introduzindo uma certa ordem na disposição das imagens. A configuração final é composta por $p(p-1)/2$ gráficos, podendo esse número ser muito elevado se p é grande, facto que levanta, certamente, alguma confusão no momento de interpretar.

O mapa com os "rabiscos do desenhador" que consta da figura 3.1 foi elaborado para o Exemplo referido no capítulo 2., e visualiza as relações que se estabelecem entre as variáveis. Destacam-se as ligações lineares entre a CPPNB e ENRG em sentido directo (os países com maior capitação são igualmente os maiores consumidores de energia) e entre PURB e PIBA em sentido inverso (naturalmente o fenómeno da urbanização diminui relativamente a dedicação à agricultura e o produto aí originado quando não é acompanhado de aumentos de produtividade que invertam o processo).

Estabelecem-se, igualmente, relações lineares entre as variáveis PURB e EPVN e PURB e HMED. As duas primeiras relacionam-se de modo directo enquanto as duas últimas o fazem em sentido inverso (a urbanização permite, em linhas gerais, melhorar as condições de saúde através da centralização dos serviços de cuidados primários e disponibilizando mais médicos para o atendimento dos doentes). Em consonância EPVN e HMED e PIBA e HMED são dois pares de variáveis que apresentam correlação, respectivamente, negativa e positiva.

FIGURA 3.1 — Os rabiscos do desenhador para 30 paises de Africa e Europa



PERFIS

Uma forma natural de representar cada observação *p*-dimensional é usar *p* barras ou colunas verticais. As barras são colocadas ao lado umas das outras com as alturas respectivamente proporcionais ao valor das *p* variáveis sendo construído um diagrama de barras por cada observação. O cimo das diferentes barras em cada diagrama pode, em alternativa, ser ligado por uma linha poligonal de modo a dar uma aparência de continuidade ao perfil.

Esta técnica, assim como as seguintes, é apropriada para representações no espaço dos indivíduos mas sofrem de dois inconvenientes: primeiro, revelam uma grande dependência da ordem por que foram dispostas as variáveis e segundo, são de difícil leitura quando *p* é elevado.

As figuras 3.2 e 3.3 mostram os perfis sob a aparência, respectivamente, de diagrama de barras e de linha poligonal, para o conjunto de dados do Exemplo. As variáveis foram transformadas porque eram medidas em unidades diferentes e originavam escalas muito díspares. A transformação escolhida, proposta por Chambers, Cleveland, Kleiner e Tukey(1983), obriga a que os novos valores se situem no intervalo [0,1]:

FIGURAS 3.2 E 3.3 – Legenda

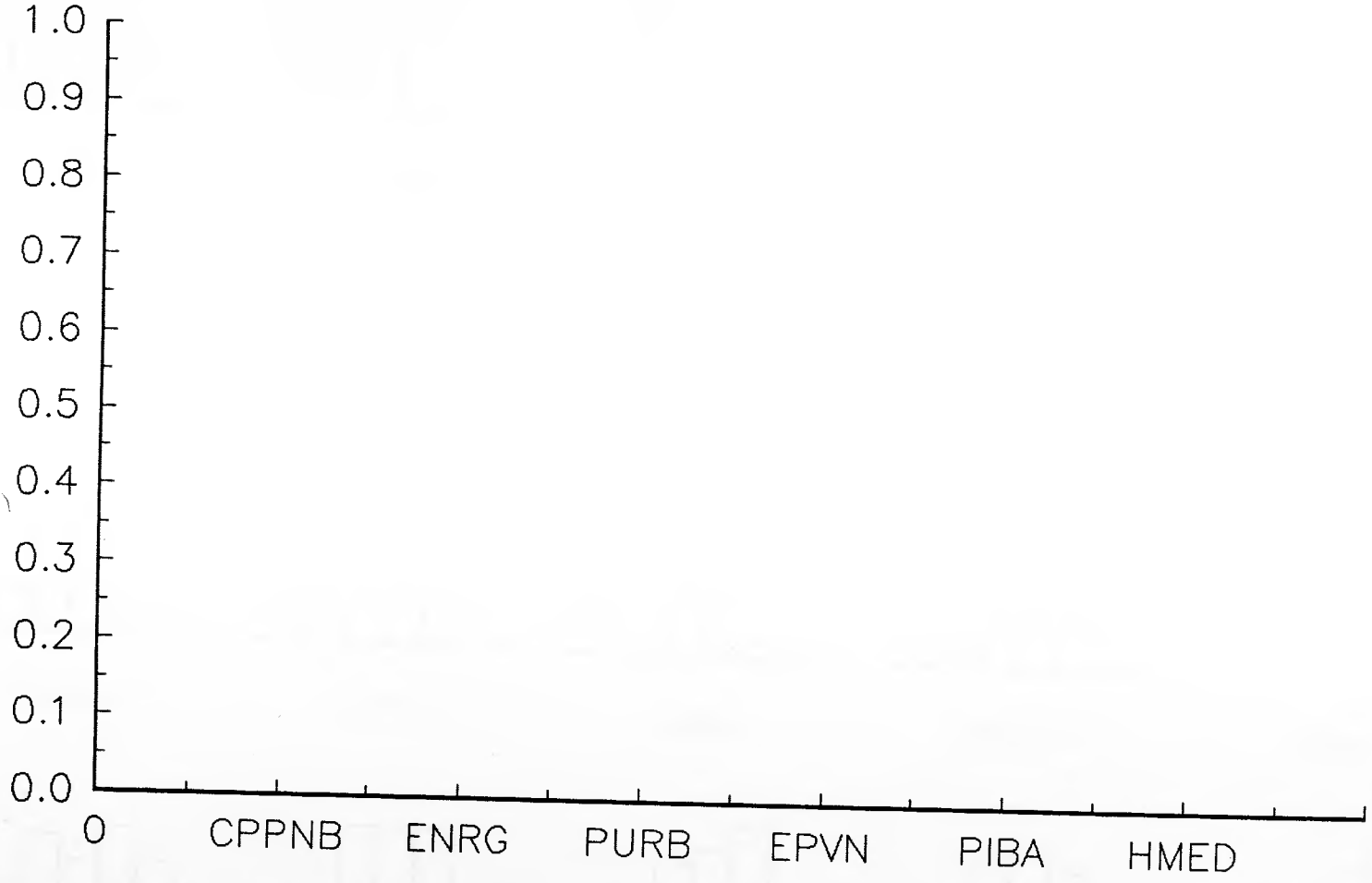


FIGURA 3.2 – *Perfis, em diagramas de barras, para 30 paises de Africa e Europa*

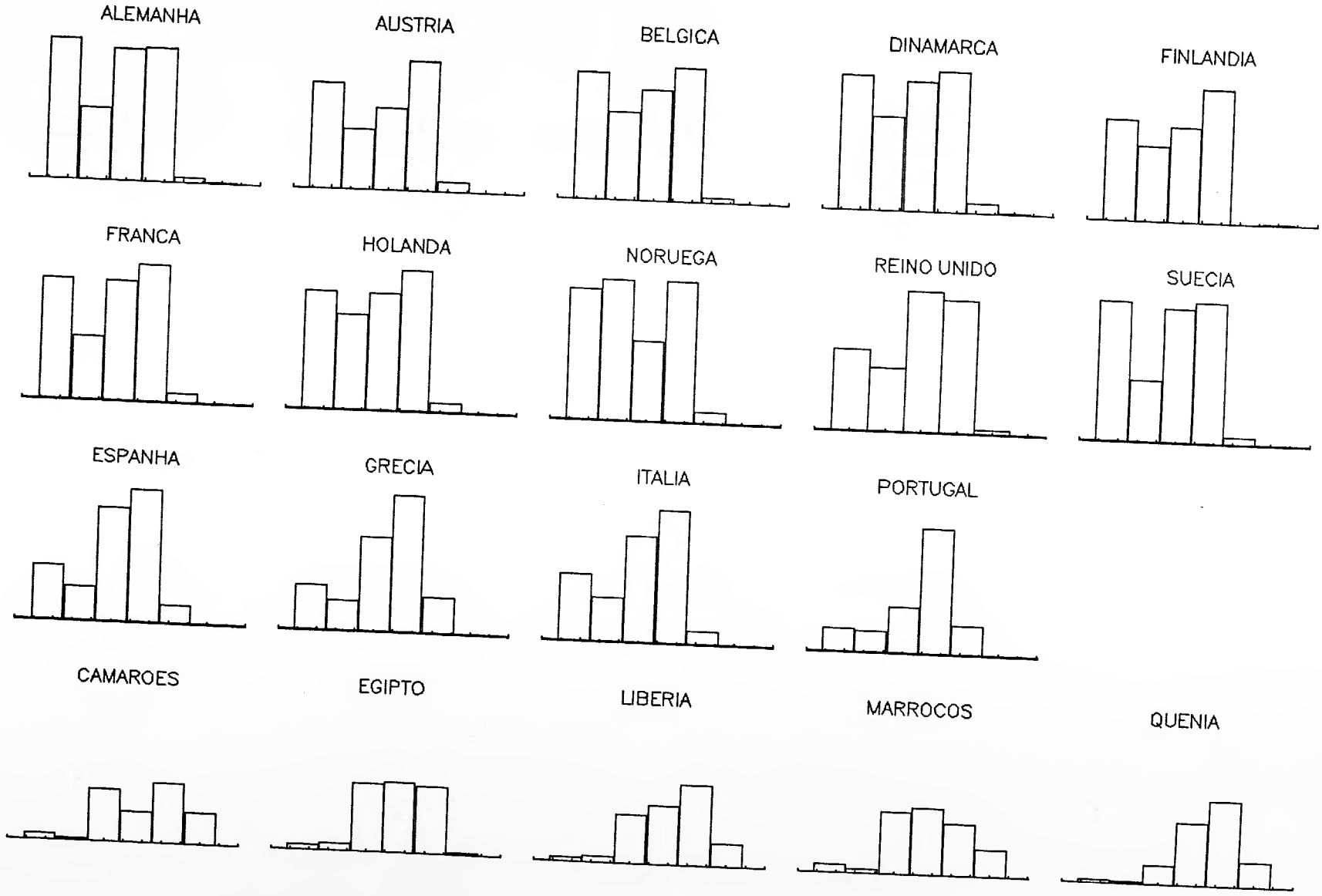


FIGURA 3.2(continuacao)

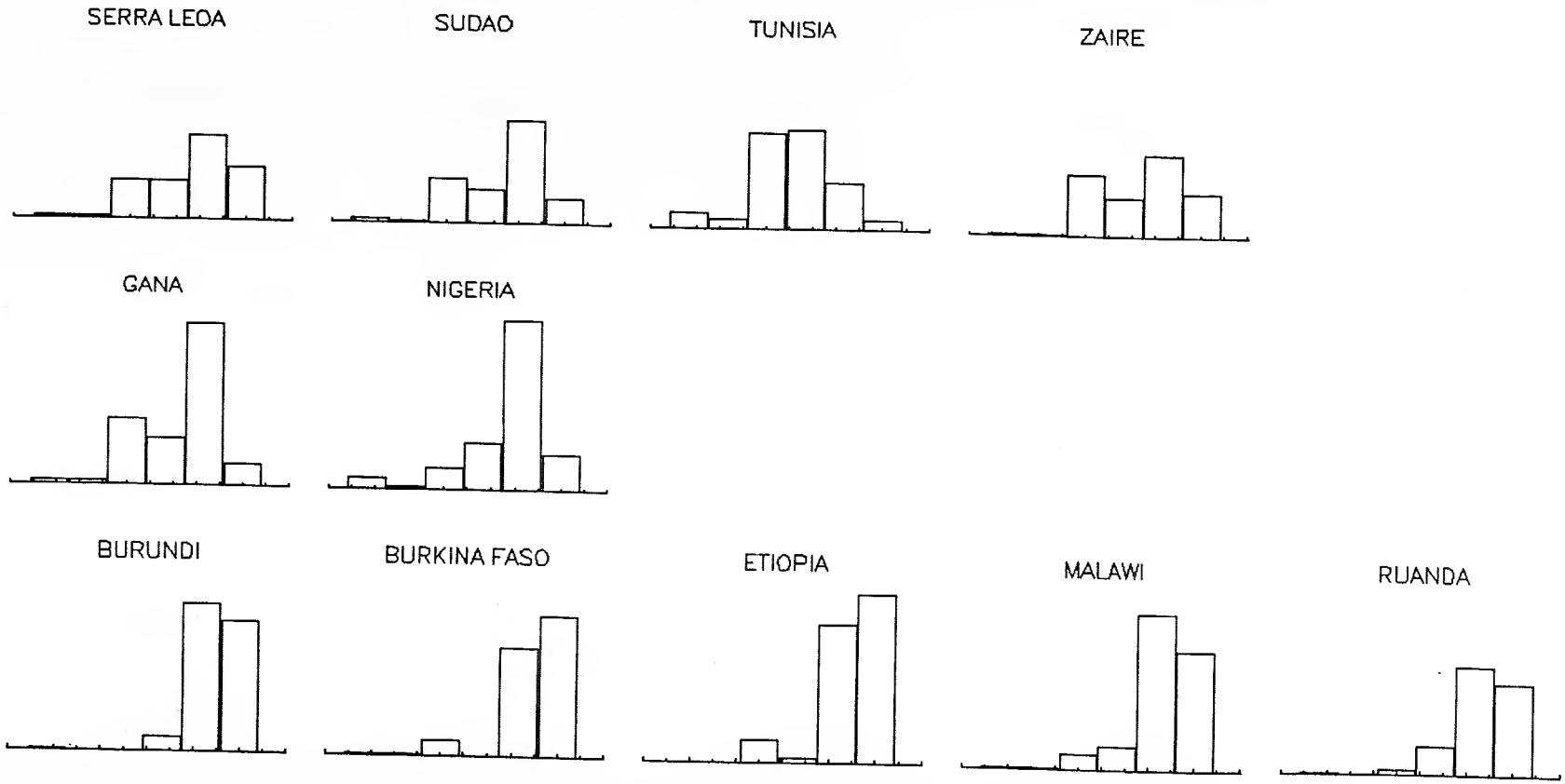


FIGURA 3.3 – Perfis, em linha poligonal, para 30 paises de Africa e Europa

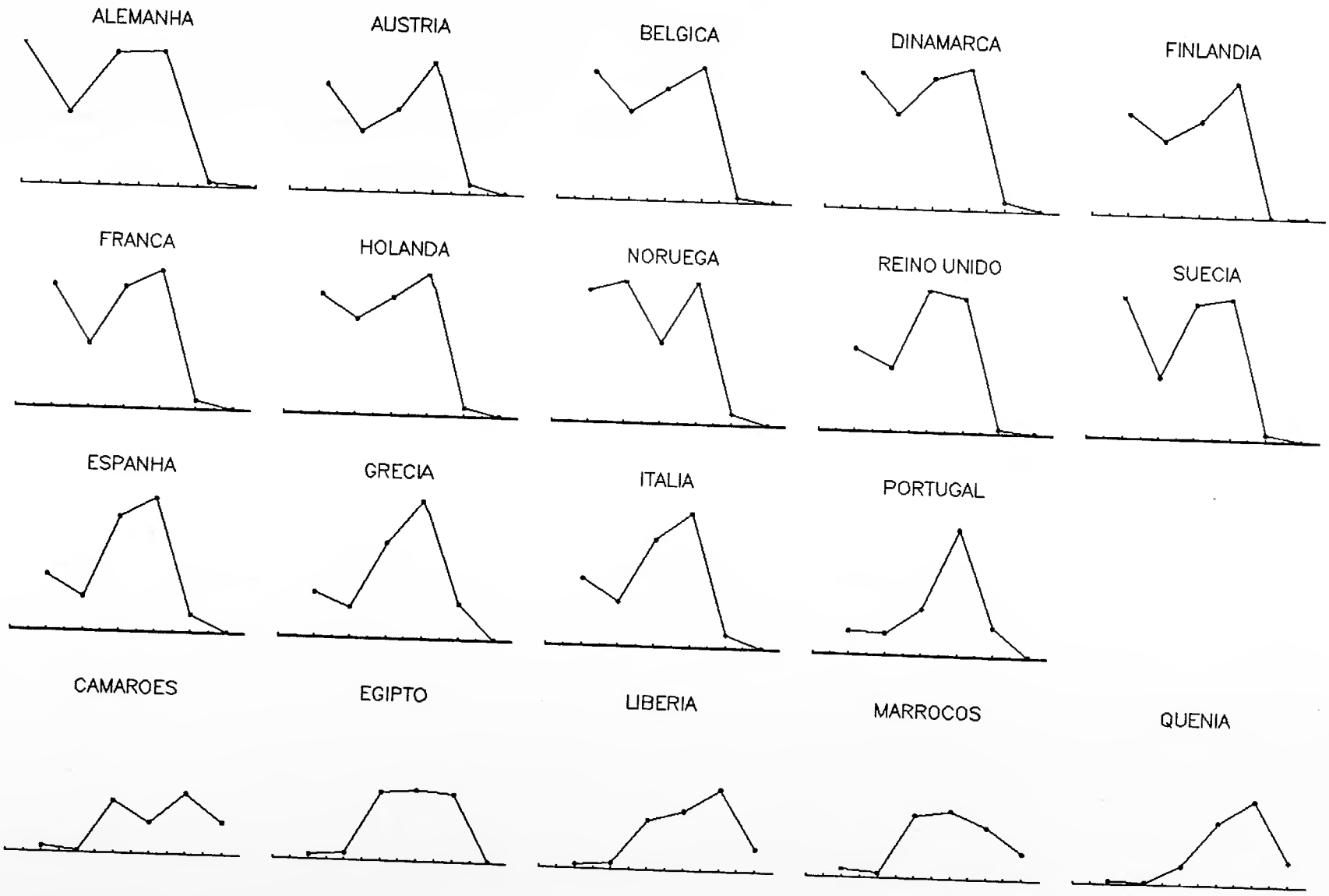
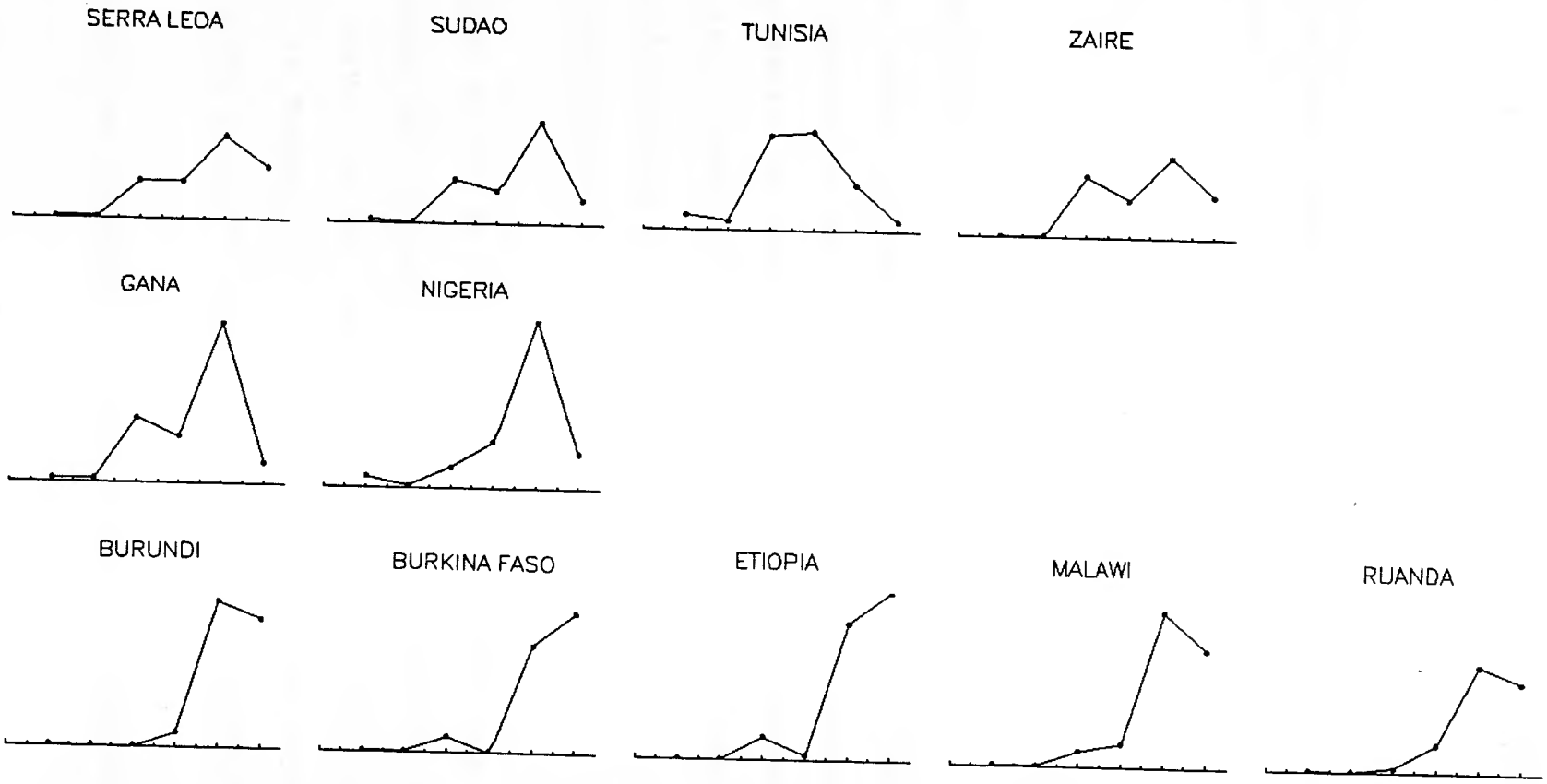


FIGURA 3.3(continuacao)



$$(3.1) \qquad x^{*}_{ij} = \frac{x_{ij} - \min_i x_{ij}}{\max_i x_{ij} - \min_i x_{ij}} \quad , \ (i=1,\ldots,n; j=1,\ldots,p).$$

Para variáveis que assumem valores negativos , a transformação proposta é:

$$(3.2) \qquad x^{*}_{ij} = x_{ij} / \max_i |x_{ij}| .$$

Olhando para as figuras 3.1 e 3.2 confirma-se a semelhança de perfil, entre os países desenvolvidos da Europa do Norte com altas capitações de PNB, elevado consumo de energia, grande parte da população em cidades e por isso baixa expressão do produto agrícola. As boas condições sociais destes países determinam que as variáveis EPVN e HMED sejam, respectivamente, significativa e não significativa.

Os países do Sul da Europa assemelham-se dado que revelam valores inferiores aos do grupo anterior nas variáveis CPPNB e ENRG. O Reino Unido, se bem que seja possível integrá-lo no grupo dos países da Europa do Norte, encontra-se numa posição intermédia entre estes e os do Sul devido à menor capitação de PNB.

Em África as condições económico-sociais apresentam perfis com

comportamento inverso aos dos países da Europa: as variáveis responsáveis pelos máximos de uns perfis determinam mínimos nos outros. Apesar das características comuns distinguem-se entre os países de África três tipos de perfis. O primeiro tipo, representativo da maior parte dos países retratados: Camarões, Egipto, Libéria, Marrocos, Quénia, Serra Leoa, Sudão, Tunísia e Zaire, expressa-se por reduzidas CPPNB e ENRG e baixos HMED e PIBA. O segundo tipo que corresponde ao Gana e à Nigéria assume grandes valores de PIBA. Finalmente o terceiro grupo de países: Burundi, Burkina Faso, Etiópia, Malawi e Ruanda caracteriza-se por elevados valores de PIBA e HMED.

ESTRELAS

As estrelas ou polígonos representam cada variável como um ponto marcado num raio que parte do centro de um círculo. Os raios são tantos quantas as variáveis, encontram-se igualmente espaçados e têm em comum a origem que é o centro do círculo. Os *p* pontos são ligados por uma linha poligonal que adquire a forma de estrela. Cada polígono representa uma observação multivariada.

As estrelas dispõem os valores segundo coordenadas polares em oposição aos perfis que o fazem em coordenadas cartesianas. Se alguma variável apresentar valores negativos deve ser transformada, por adição de uma constante, de forma a representar

unicamente valores positivos. Neste caso concreto os perfis superam as estrelas porque os valores negativos são permitidos, conduzindo a perfis que passam abaixo do eixo das abcissas.

Chambers, Cleveland, Kleiner e Tukey(1983) enunciam as etapas que a construção de uma estrela percorre:

1. As variáveis são transformadas com (3.1) para que os valores se situem no intervalo [0,1].

2. O j-ésimo raio forma um ângulo,

$$\theta_j = 2\pi(j-1)/p \ , \tag{3.3}$$

a partir da horizontal.

3. Impõe-se uma medida máxima para os raios, seja R e ligam-se os pontos de 1 até p e de novo até 1 de modo a fechar a estrela,

$$P_{ij} = (x^*_{ij} R \cos \theta_j, x^*_{ij} R \text{ sen } \theta_j) \ . \tag{3.4}$$

As estrelas podem ser adaptadas para destacar a relação entre os valores assumidos pelas variáveis e as respectivas médias em cada observação. Neste caso usa-se o meio de cada raio para representar a média da variável. Os valores de cada variável

encontram-se nos lados do polígono e não nos raios. Este tipo de estrela designa-se, neste ensaio, por modelo 2 e o anterior por modelo 1.

A representação em estrela é mais sofisticada que os perfis e o facto de os pontos de partida corresponderem aos pontos de chegada facilita a comparação entre sujeitos. A desvantagem das estrelas surge quando o número de variáveis é grande e as figuras podem ter um aspecto esquisito e confuso.

A figura 3.4 contém as estrelas modelo 1 para o Exemplo. Uma vez mais, Alemanha, Áustria, Bélgica, Dinamarca, Finlândia, França, Holanda, Noruega, Suécia e Reino Unido formam um grupo de características comuns com grandes estrelas mas imperfeitas: as pontas viradas para baixo, que correspondem a PIBA e HMED, não existem. Espanha, Itália e Grécia são representados por estrelas que têm o mesmo formato das estrelas dos países anteriores mas de tamanho mais reduzido. A estrela portuguesa não foi emparceirada porque, se bem que aparente semelhanças com a grega, é mais pequena.

Os países de África agrupam-se em três conjuntos para os quais as respectivas estrelas possuem em comum a pequena dimensão e se distinguem nos aspectos que se enumeram de seguida. Camarões, Zaire, Marrocos, Serra Leoa e Tunísia são representados pelas estrelas mais parecidas com o símbolo estilizado do astro porque

FIGURA 3.4 – Legenda

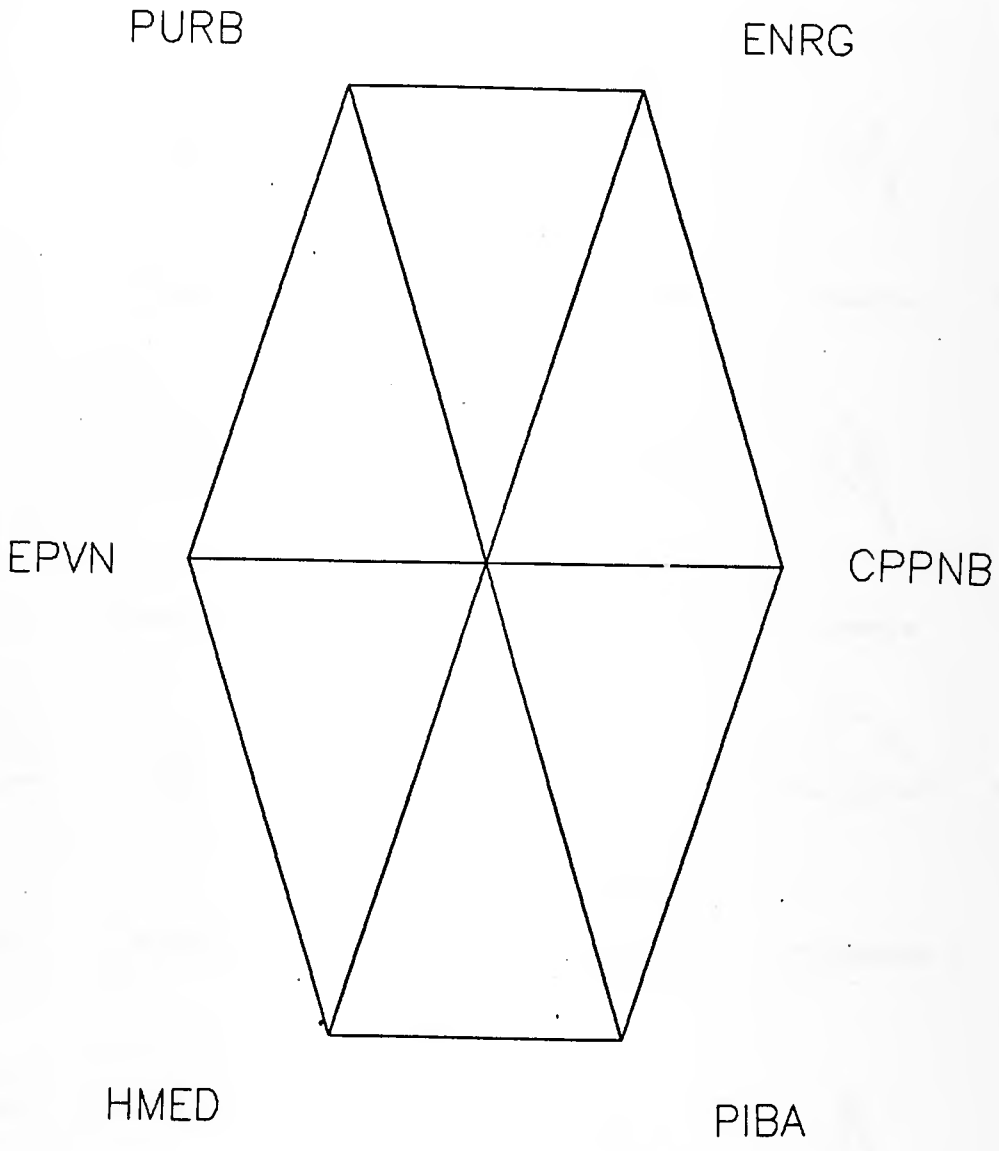
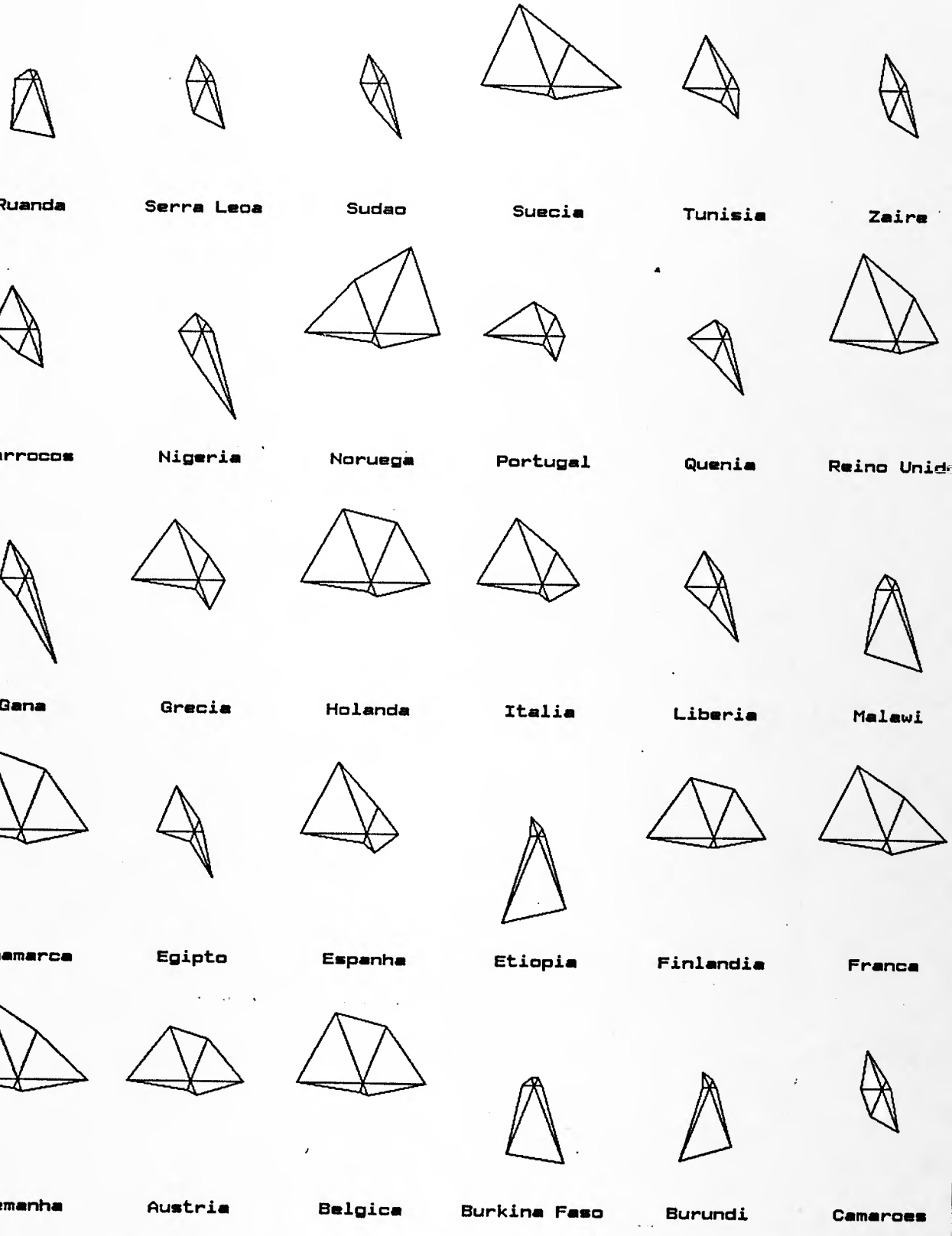


FIGURA 3.4 – Estrelas (modelo 1)
para 30 paises de Africa e Europa



as pontas são todas visíveis. Gana, Egipto, Libéria, Nigéria, Quénia e Sudão apresentam estrelas em que a ponta virada para baixo à direita, relativa à variável PIBA, é muito pronunciada. Finalmente, as estrelas dos países da cauda do desenvolvimento de entre os que aqui são tratados, Burkina Faso, Burundi, Etiópia, Malawi e Ruanda apenas têm as pontas viradas para baixo. Mais uma vez se verifica possuirem estes países situações económico-sociais inversas das dos países da Europa do Norte.

A figura 3.5 retrata as estrelas do 2º modelo para os mesmos países. A leitura é idêntica à do gráfico da figura 3.4 sendo possível, adicionalmente verificar que, por exemplo, para os países da Europa do Norte as variáveis situadas nos raios virados para baixo (HMED e PIBA) assumem valores bastante inferiores à média e as restantes variáveis o contrário. Para o grupo de países africanos Burkina, Burundi, Etiópia, Malawi e Ruanda, como seria de esperar, as variáveis situadas nos raios virados para baixo assumem valores superiores à média e as outras variáveis têm comportamento inverso.

GLYPHS

Anderson(1960) propôs uma técnica que apelidou de ***glyphs*** na qual cada observação é representada por um círculo e cada variável por um raio originado no círculo. Os raios são desenhados a partir da

FIGURA 3.5 – Legenda

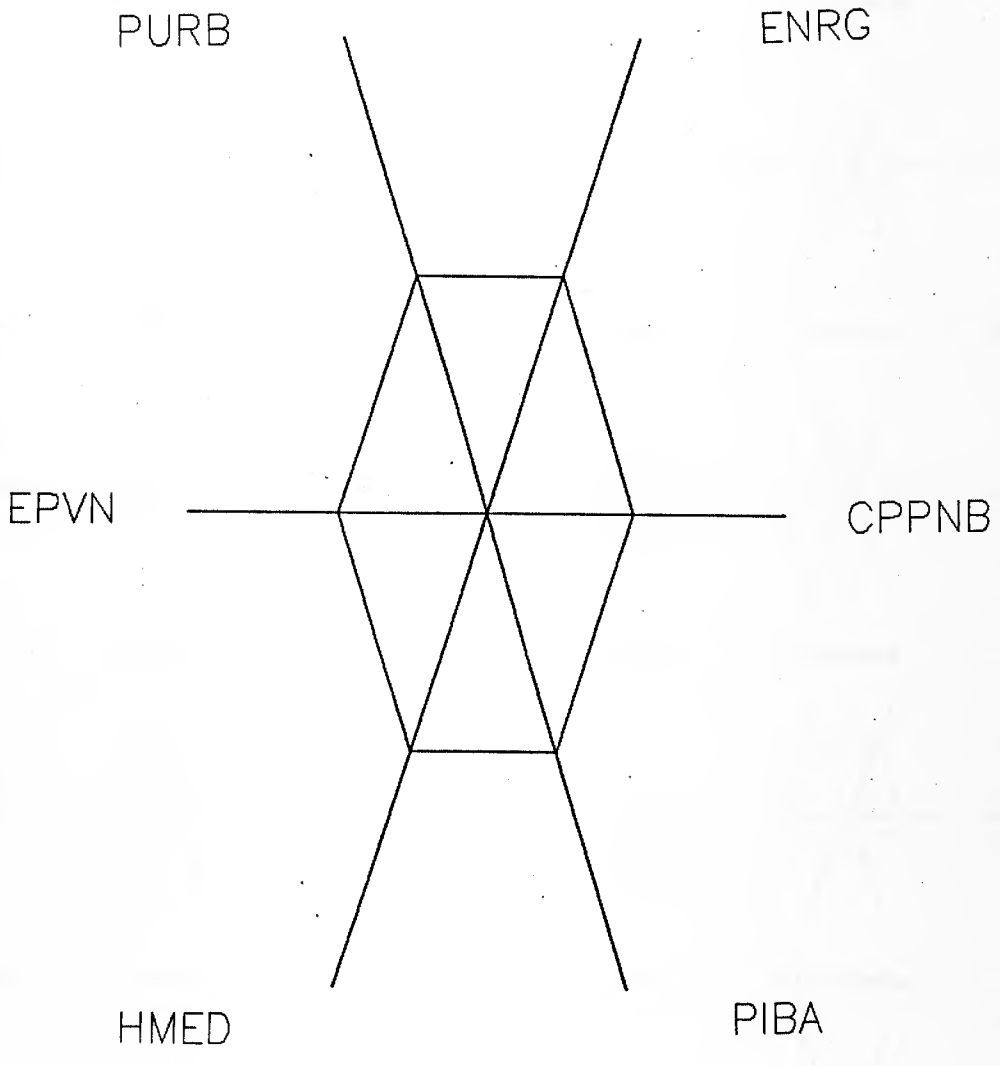
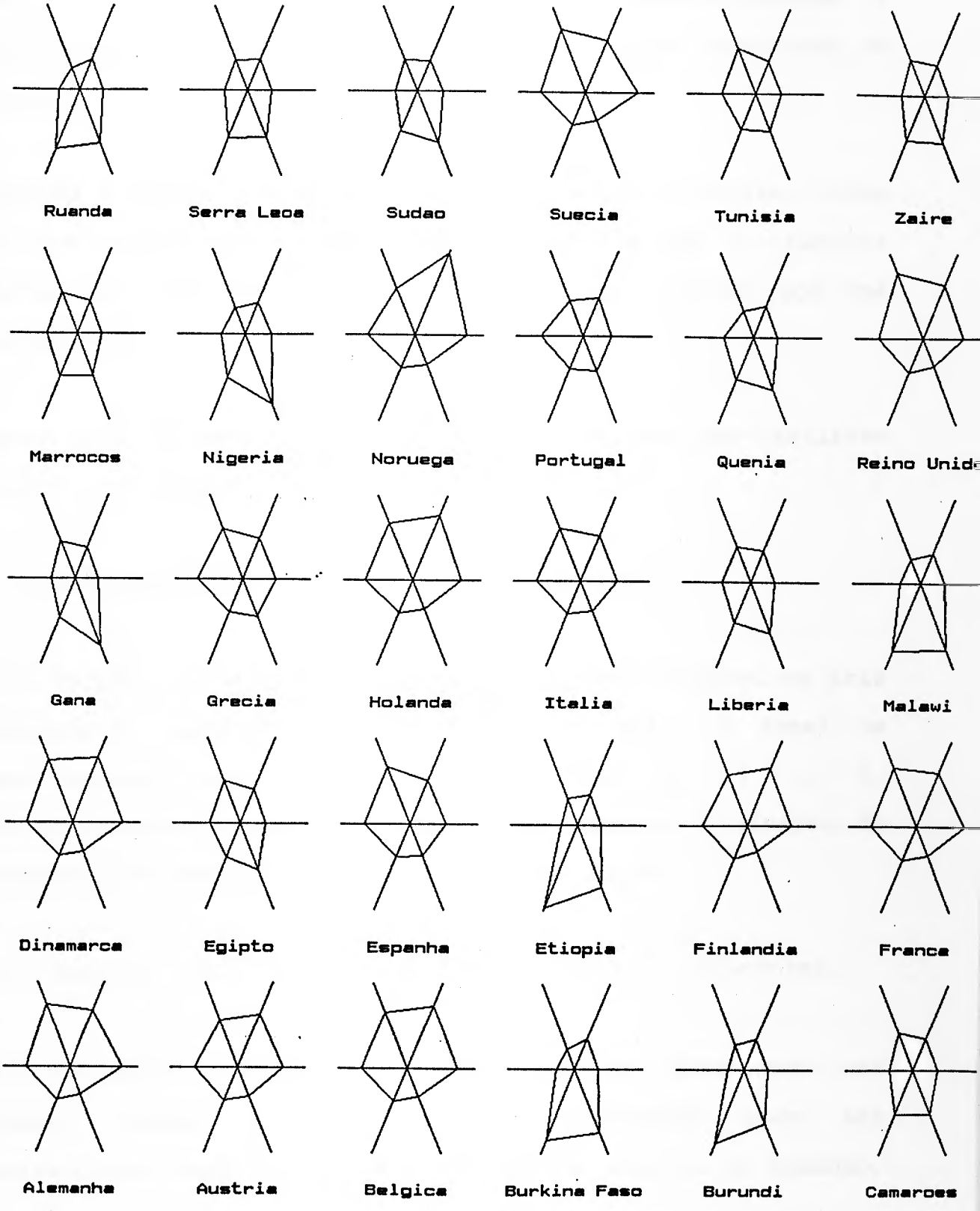


FIGURA 3.5 – Estrelas (modelo 2)
para 30 paises de Africa e Europa



circunferência que delimita o círculo e dirigidos para fora dele. A posição e o comprimento do raio indica, respectivamente, a variável que está a ser representada e o valor observado da variável.

Os *glyphs* diferem das estrelas em dois aspectos. Primeiro, todos os raios partem da extremidade do círculo e não do centro; segundo, os pontos marcados nos raios não são ligados por uma linha poligonal.

O mesmo autor forneceu, ainda, indicações práticas que facilitam a leitura dos *glyphs*:

i) Não representar mais de sete variáveis.

ii) Dividir o intervalo de variação de cada variável em três categorias, nomeadamente baixa, média e alta, e fazer os comprimentos dos raios correspondentes 0, 1 e 2, respectivamente. Como alternativa pode usar-se a divisão do intervalo de variação em quartis ou mesmo decis.

iii) Associar variáveis correlacionadas em raios adjacentes.

iv) Se dois ou mais tipos divergentes são observados (por exemplo homens e mulheres) essa informação pode ser representada distinguindo os círculos (por exemplo no tamanho,

na cor).

v) É possível desenhar os *glyphs* num conjunto de eixos cartesiano e nestes representar duas variáveis retirando dois raios aos *glyphs*.

As desvantagens desta técnica são a dependência face à ordenação das variáveis e o surgimento de figuras confusas quando o número de variáveis é grande.

A figura 3.6 ilustra os *glyphs* para o Exemplo. As variáveis foram divididas em quatro categorias por ordem crescente, a que correspondem comprimentos de, respectivamente 0, 1, 2 e 3 unidades de eixo. Os países de África são representados pelos círculos de menor diâmetro enquanto que os países da Europa são representados pelos de maior diâmetro.

Os países da Europa do Norte têm *glyphs* semelhantes caracterizados por possuirem quatro grandes raios, dois do lado direito e dois do lado esquerdo do círculo o que confere aos símbolos um aspecto equilibrado.

Espanha, Grécia, Itália e Portugal constituem um outro grupo mais heterogéneo a que se pode juntar a Tunísia, o Egipto e Marrocos. Os *glyphs* deste grupo transmitem, igualmente, uma ideia de equilíbrio em que o lado esquerdo e o lado direito se

FIGURA 3.6 — Legenda

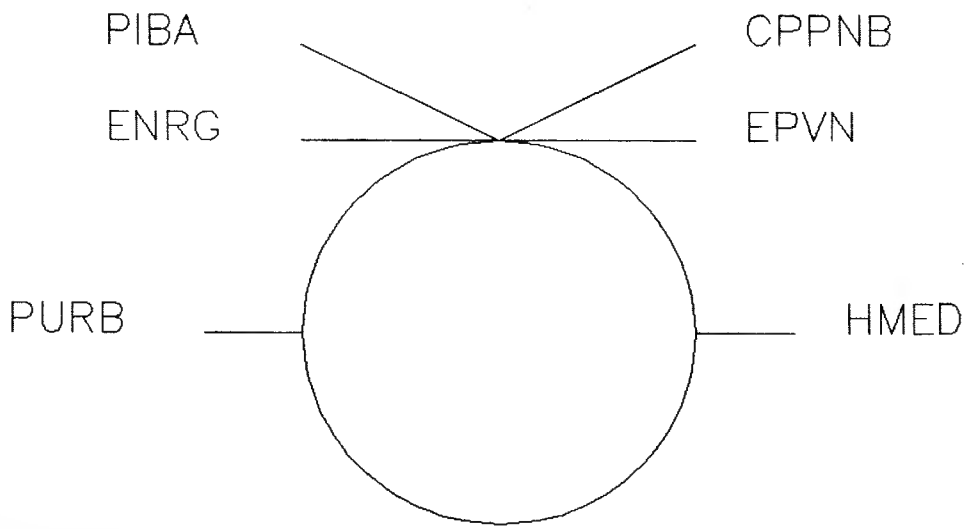


FIGURA 3.6 – GLYPHS PARA 30 PAISES DE AFRICA E EUROPA

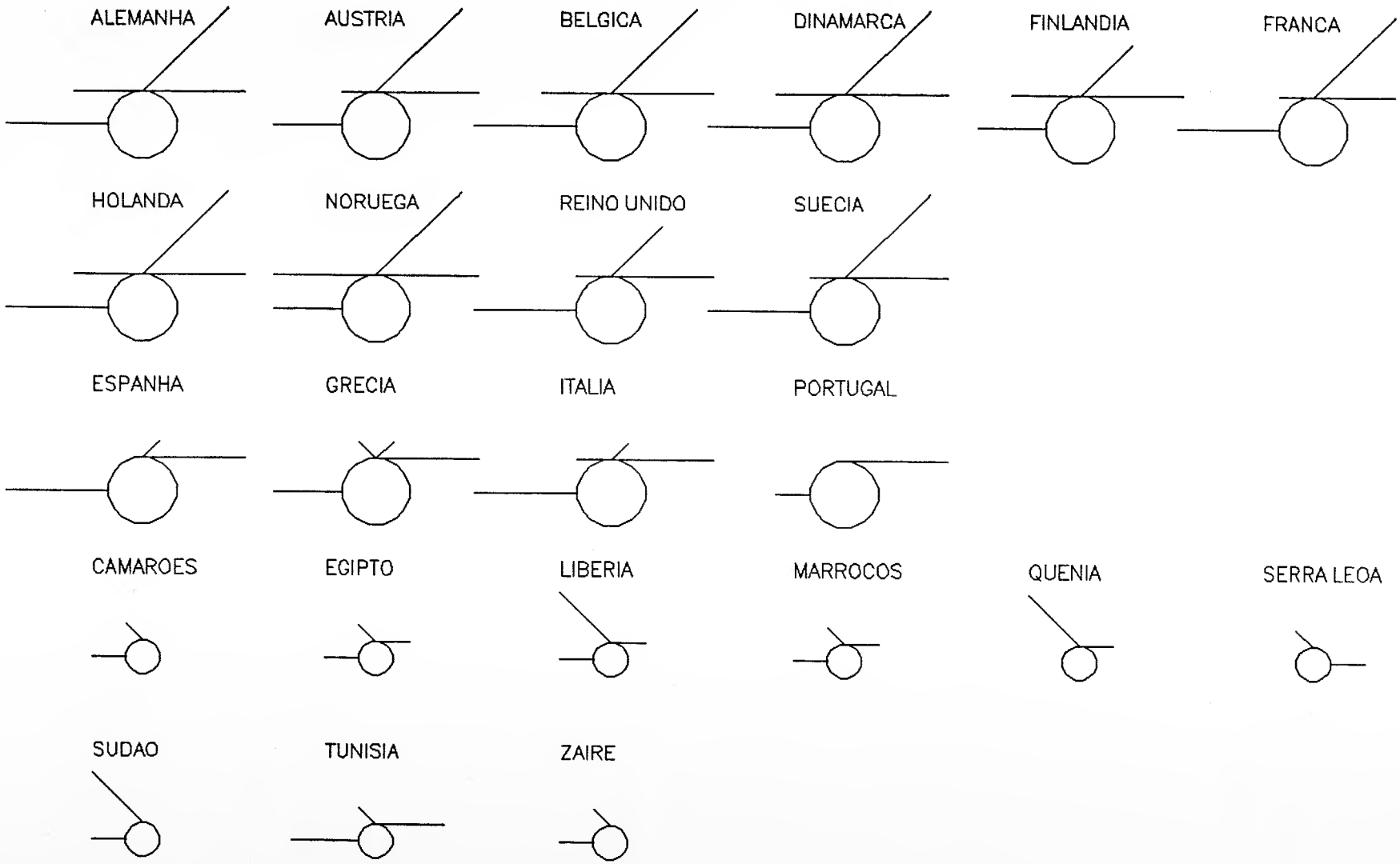
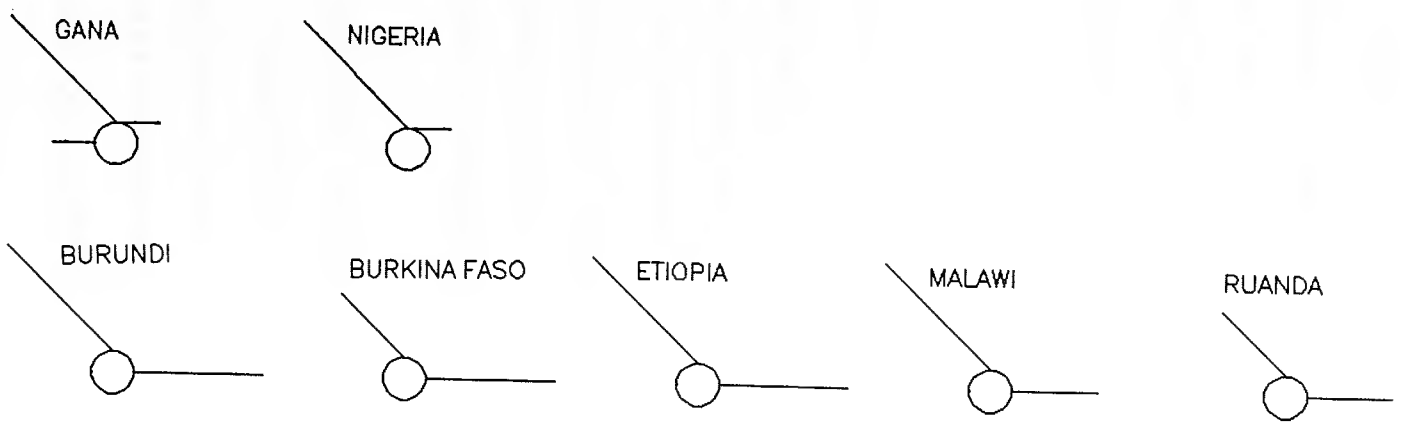


FIGURA 3.6 (continuacao)



contrabalançam.

Com os restantes países africanos formam-se três grupos. Camarões, Sudão e Zaire constituem o primeiro cujos *glyphs* têm apenas dois raios e virados para o lado esquerdo. Libéria, Quénia, Gana e Nigéria formam o segundo grupo em que os *glyphs* são dominados pela variável PIBA. A associação dos países que aparecem na última linha do gráfico caracteriza-se pela grande expressão das variáveis PIBA e HMED dando aos respectivos *glyphs* a aparência de ângulo obtuso.

CAIXAS

Em alternativa à representação de dados através de linhas e círculos surge a ideia de associar o valor de uma variável à medida da característica de um objecto familiar. A versão mais simples desta abordagem é enunciada por Hartigan(1975) que recomendou o uso de caixas para representar dados multivariados associando as variáveis às dimensões da caixa.

Se o número de variáveis, p, é igual a 3, existe correspondência entre o comprimento, largura e altura da caixa e as variáveis. Para $p>3$ cada lado da caixa é dividido em segmentos e é, assim, usado para representar mais do que uma variável. Nestes casos, as variáveis correlacionadas devem ser colocadas em segmentos do

mesmo lado uma vez que se torna mais fácil aperceber as diferenças entre os comprimentos dos lados do que entre os segmentos de um lado específico. Agrupar variáveis correlacionadas num dos lados da caixa reduz a variação entre os comprimentos dos segmentos desse lado.

As caixas partilham as desvantagens dos métodos anteriores: são sensíveis à ordenação das variáveis e tornam-se pouco úteis quando o número de variáveis é superior a 8.

AS CARAS DE CHERNOFF

O pioneiro da linha de pensamento que associa variáveis a características de objectos familiares foi Chernoff. Em 1973, (Chernoff,1973) propôs uma versão notável, embora complexa, associando o valor de cada variável a um aspecto da face humana como, por exemplo, a dimensão e curvatura da boca, o tamanho dos olhos, o desenho das sobrancelhas, etc.. Uma caricatura do rosto é elaborada por cada observação multivariada.

Dois motivos determinaram a escolha das caras. Primeiro, qualquer observador reconhece e responde, com facilidade, a mudanças nas expressões faciais extraindo visualmente o que é importante e abstraindo do que é insignificante. Segundo, é possível atribuir significado psicológico às características faciais, como seja, um

sorriso representar uma variável de sucesso ou os olhos estarem associados a uma variável que revele astúcia.

O programa escrito por Chernoff utilizava, no máximo, 18 variáveis. Bruckner(1978) introduziu mais duas características, a largura do nariz e o diâmetro dos ouvidos, enquanto Flury e Riedwyl(1981) alargaram o método a 36 variáveis. Tal foi conseguido transformando características simétricas como nariz, boca, cabelo, em assimétricas, ou seja, possibilitando que a parte esquerda, por exemplo, da boca tenha um aspecto diferente da parte direita.

Quando são observadas poucas variáveis algumas características ficam por atribuir, o que significa permanecerem constantes em todas as caras e neste caso devem escolher-se as menos notadas. Se determinadas variáveis estão relacionadas é útil associá-las a uma área específica do rosto, por exemplo, aos olhos e às sobrancelhas.

As variáveis mais importantes devem ser atribuídas às características mais relevantes do rosto. Mas esta escolha é subjectiva: diferentes observadores utilizam diferentes características para julgar semelhanças. Uns concentram-se nos olhos, outros no queixo, etc. Pode tentar ultrapassar-se este problema produzindo vários conjuntos de rostos cada um envolvendo diferentes associações variáveis-características. À volta da

subjectividade divagaram autores como Everitt(1978) e Bruckner(1978). Para o primeiro este aspecto constitui, sem dúvida, um inconveniente enquanto que, para o segundo, é uma vantagem na medida em que obriga o investigador a ser cauteloso nas conclusões. Chernoff e Rizvi(1975) estimaram existir, num exercício de classificação, uma taxa de erro de cerca de 25% causada por permutações aleatórias das associações variáveis-características da face.

Um vasto campo de aplicações desta técnica é apresentado por vários autores em Wang(1978).

A representação do Exemplo através das caras de Chernoff consta da figura 3.7. A dualidade Europa-África continua a ser visível nesta representação assim como as diferenças existentes dentro de cada continente. Os países do hemisfério norte caracterizam-se por possuirem uma fisionomia aberta com nariz pequeno enquanto que os países do hemisfério sul aparentam rostos mirrados de nariz comprido. A largura da testa varia em sentido directo com o estádio de desenvolvimento dos países e a dimensão do nariz em sentido inverso, traduzindo-se deste modo as diferenças entre os países intra Europa e intra África. Entre os países europeus distinguem-se os agrupamentos habituais: Europa do Norte e Europa do Sul. Entre os países africanos formam-se duas classes abrangendo a primeira o Burkina Faso, o Burundi, a Etiópia, o Malawi e o Ruanda e a segunda classe os restantes países.

FIGURA 3.7 – Legenda

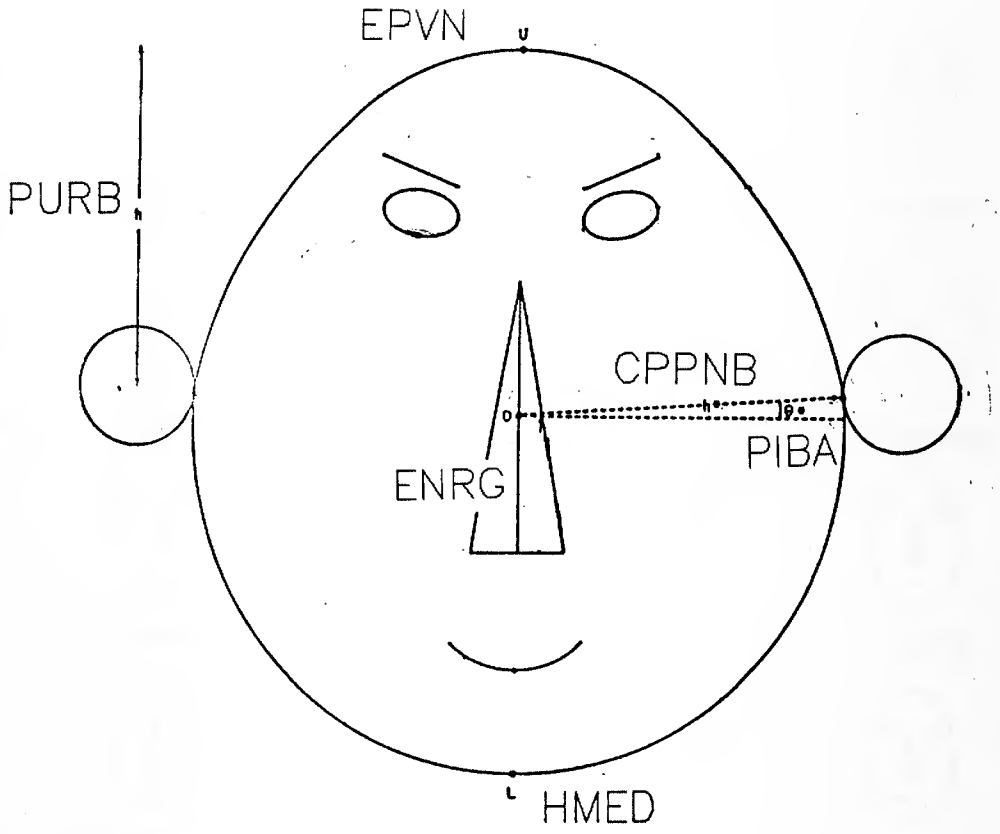
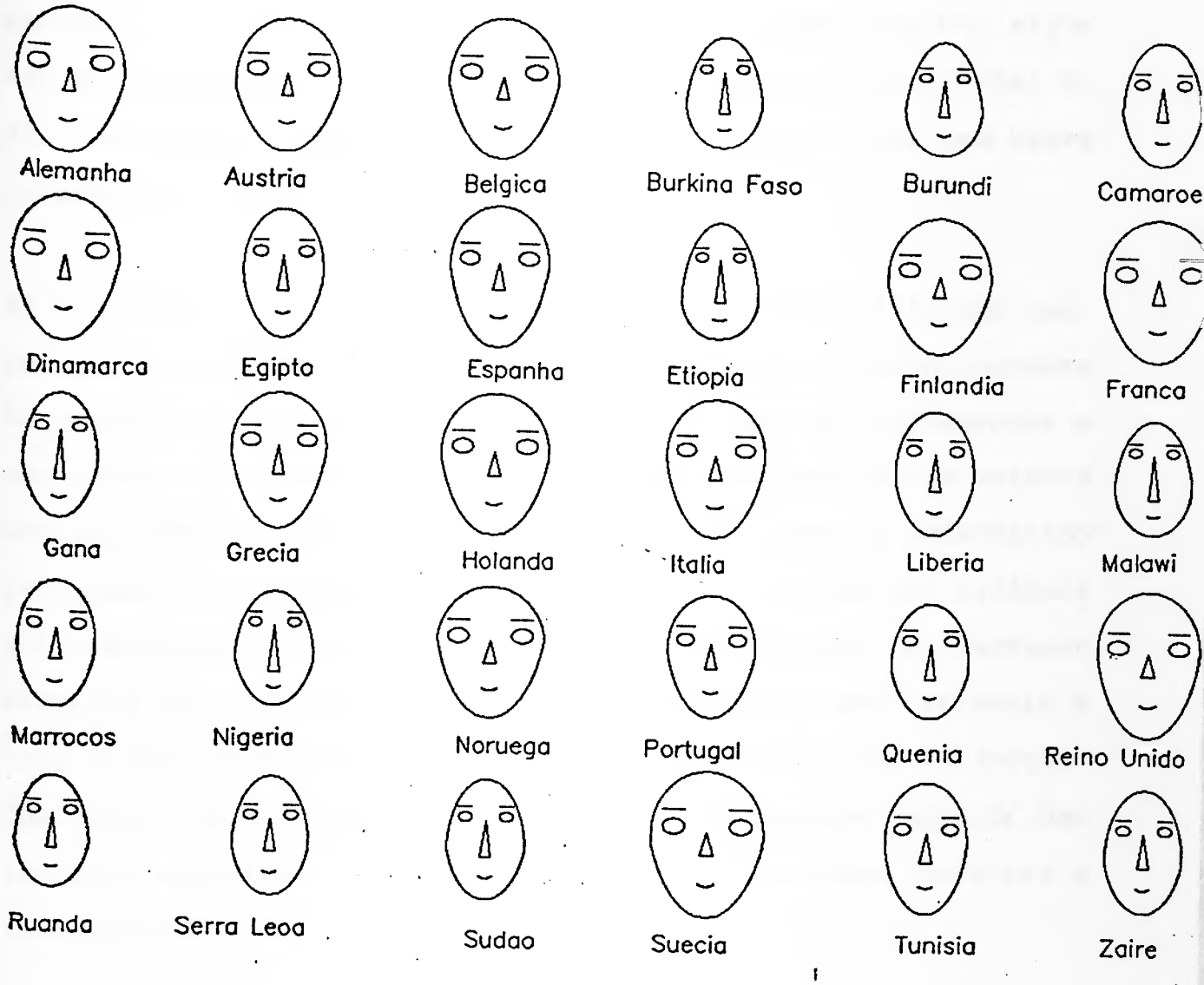


FIGURA 3.7 – Caras de Chernoff
para 30 paises de Africa e Europa



ÁRVORES

Como representações intermédias entre as simples caixas e os mais complexos rostos, Kleiner e Hartigan(1981) sugeriram árvores e castelos. A técnica descrita por estes autores envolve algum estudo preliminar, no âmbito da análise de agrupamentos, antes de ser elaborado o gráfico. Por esse motivo apenas se faz uma breve referência ao assunto.

As variáveis são atribuídas a ramos de uma árvore estilizada que, por seu turno, são ligados a troncos secundários e eventualmente ao tronco principal. Cada observação multivariada corresponde a uma árvore e os ramos têm comprimentos determinados pelos valores das variáveis. O comprimento de cada ramo interno é determinado pela média de todos os ramos que suporta. Em vez de atribuir arbitrariamente as variáveis aos ramos, Kleiner e Hartigan efectuam uma análise de agrupamento hierárquico das variáveis e usam o dendograma resultante como base para desenhar a árvore. Com isto, os autores tentam obviar a desvantagem sentida nas técnicas anteriores em que a sequência das variáveis modifica a representação.

As árvores são apropriadas para dados nos quais existam agrupamentos de variáveis altamente correlacionadas. As variáveis

percententes ao mesmo agrupamento estarão situadas no mesmo lado da árvore.

Este procedimento pode ser modificado para obter figuras que se assemelham a castelos fazendo o ângulo entre ramos igual a zero e desenhando os ramos uns ao lado dos outros. A interpretação dos castelos é idêntica à leitura das árvores.

3.2 AS CURVAS DE ANDREWS

A representação de dados multivariados através de funções matemáticas foi proposta por Andrews(1972) e surgiu como alternativa à versão simbólica referida no ponto anterior. Aquele autor escolheu a função harmónica de Fourier devido às convenientes propriedades que possui e, até agora, é esta função que, tem tido mais utilizações. A técnica funcional insere-se na abordagem directa do espaço dos indivíduos.

Cada observação multivariada é transformada numa função $f_{\mathbf{x}}(t)$ de uma só variável t que se expressa como combinação linear de funções ortogonais $f(t)$. Os coeficientes da combinação linear não são mais que os valores observados das variáveis $X_1, \ldots, X_p$.

Andrews (1972) sugeriu que $f_{\mathbf{x}}(t)$ fosse a função de Fourier e,

assim, as funções $f(t)$ são {sen kt} e {cos kt} com k inteiro. Contendo o vector $\mathbf{x} = (x_1,\ldots,x_p)$ os valores das variáveis para uma observação, a função $f_{\mathbf{x}}(t)$ correspondente é definida pela expressão,

$$(3.1) \qquad f_{\mathbf{x}}(t) = x_1 2^{-1/2} + x_2 \operatorname{sen} t + x_3 \cos t + x_4 \operatorname{sen} 2t + x_5 \cos 2t + \ldots, \quad -\pi < t < \pi.$$

Para o conjunto das n observações obtêm-se n funções, $f_{\mathbf{x}_1}(t),\ldots,f_{\mathbf{x}_n}(t)$, que são representadas simultaneamente num referencial cartesiano em que as ordenadas correspondem aos valores de $f_{\mathbf{x}}(t)$ e as abcissas aos valores possíveis de t ($-\pi<t<\pi$ ou, então, substitui-se t por $2\pi t$ e usa-se o intervalo $0<t<1$). As observações multivariadas, associadas a n pontos num espaço p-dimensional, reduzem-se a n curvas num espaço bidimensional.

A função $f_{\mathbf{x}}(t)$ possui três interessantes propriedades:

i) A representação em termos de funções $f_{\mathbf{x}}(t)$ preserva as distâncias euclideanas: a distância quadrática entre o par de funções $f_{\mathbf{x}_l}(t)$ e $f_{\mathbf{x}_k}(t)$ é proporcional à distância quadrática euclideana entre as observações $\mathbf{x}_l$ e $\mathbf{x}_k$ de coordenadas, respectivamente, $\mathbf{x}_l=(x_{l1},\ldots,x_{lp})$ e $\mathbf{x}_k=(x_{k1},\ldots,x_{kp})$,

$$(3.2)^* \qquad \|f_{\mathbf{x}_l} - f_{\mathbf{x}_k}\|^2 = \int_{-\pi}^{\pi} [f_{\mathbf{x}_l} - f_{\mathbf{x}_k}]^2 dt = \pi \left[\sum_{j=1}^{p} (x_{lj} - x_{kj})^2 \right].$$

Em consequência, pontos que se encontrem próximos no espaço p-dimensional das observações são representados por curvas que, no gráfico, estão desenhadas junto umas das outras para todos os valores de t; pontos distantes são representados por curvas que se situam afastadas para pelo menos alguns valores de t.

ii) A transformação $f_{\mathbf{x}}(t)$ preserva a média: a observação média representada pelo vector das médias das variáveis, $\bar{\mathbf{x}}$, corresponderá a uma curva que é a média das funções $f_{\mathbf{x}}(t)$, $\overline{f_{\mathbf{x}}(t)}$,

$$(3.3)^* \qquad f_{\bar{\mathbf{x}}}(t) = \overline{f_{\mathbf{x}}(t)}, \quad -\pi<t<\pi.$$

Estas propriedades permitem a identificação de agrupamentos, *outliers* e outras particularidades dos dados ou a comparação de funções individuais com a função média.

Andrews formulou, ainda, testes de significância e intervalos de confiança para as funções $f_{\mathbf{x}}(t)$ assumindo que as p variáveis não são correlacionadas. A terceira propriedade a enunciar é usada neste contexto.

iii) A transformação $f_{\mathbf{x}}(t)$ preserva a variância: se as p

variáveis têm variâncias iguais, σ^2, e não estão correlacionadas, a variância de $f_{\mathbf{x}}(t)$ para diferentes valores de t é essencialmente constante:

$$(3.4)^* \quad V[f_{\mathbf{x}}(t)] = \begin{cases} (1/2)\,\sigma^2\,p & (p \text{ ímpar}) \\ (1/2)\,\sigma^2\,[p-1+2\,\text{sen}^2(pt/2)] & (p \text{ par}). \end{cases}$$

Verifica-se que, no caso de p ser ímpar, a $V[f_{\mathbf{x}}(t)]$ se reduz à constante $(1/2)\sigma^2 p$ e, quando p é par, aquela variância assume um valor entre $(1/2)\sigma^2(p-1)$ e $(1/2)\sigma^2(p+1)$. A influência de t sobre a $V[f_{\mathbf{x}}(t)]$ é na primeira situação nula e na segunda muito ligeira, diminuindo à medida que p aumenta.

Se o conjunto de dados de que se dispõe constituir uma amostra pode ser interessante testar hipóteses ou construir intervalos de confiança para a curva representativa da média da população a partir da qual a amostra foi retirada: $f_{\mu}(t)$, sendo μ o vector das médias da população para as p variáveis.

Por vezes, identificam-se, *a priori*, certos valores de t para os quais se deseja testar a hipótese,

$$(3.5) \qquad E[f_{\mathbf{x}}(t)] = f_{\mu_0}(t),$$

avaliando o nível de significância de

$$Z = \frac{\{f_{\mathbf{x}}(t) - f_{\mu_0}(t)\}}{\sqrt{V\,[f_{\mathbf{x}}(t)]}}, \tag{3.6}$$

em que μ_0 se assume ser uma constante conhecida.

Se as p variáveis são independentes, normalmente distribuídas e têm variância comum, então Z tem distribuição normal estandardizada sob a hipótese $\mu=\mu_0$. Esta distribuição pode, pois, ser usada, se σ^2 é conhecido, para atribuir confiança à hipótese (3.5) ou para construir intervalos de confiança para μ.

Sob as mesmas hipóteses de independência das variáveis e igualdade das variâncias, Andrews estabeleceu um teste global. Se σ^2 é conhecida, a função $f_{\mathbf{x}}(t)$ situa-se com probabilidade $(1-\alpha)$, no intervalo centrado em $f_{\mu}(t)$ e amplitude constante,

$$|f_{\mathbf{x}}(t) - f_{\mu}(t)|^2 \leq (1/2)\sigma^2(p+1)\chi^2_{p,\alpha} \tag{3.7}$$

onde $\chi^2_{p,\alpha}$ designa o limiar de $100(1-\alpha)\%$ da distribuição χ^2 com p graus de liberdade, para todos os valores de t. Se μ é conhecido os *outliers* estarão fora deste intervalo e serão facilmente detectados. Um intervalo da mesma amplitude centrado em $f_{\mathbf{x}}(t)$ é uma região de confiança para $f_{\mu}(t)$, de modo que, se $f_{\mu_0}(t)$ estiver fora do intervalo há evidência contra a hipótese $\mu=\mu_0$.

As formas simples das estatísticas anteriores dependem da terceira propriedade baseada nas hipóteses de não correlação e igual variância das variáveis. Gooldchild e Vijayan(1974) alargaram o uso dos testes de significância a situações em que tais hipóteses não se verificam. Aqueles autores sublinharam que se a matriz de covariâncias da amostra é **W**, a variância de $f_{\mathbf{x}}(t)$ é dada por,

$$V\,[f_{\mathbf{x}}(t)] = \mathbf{u}^{T}\,\mathbf{W}\,\mathbf{u}, \tag{3.8}$$

com $\mathbf{u}^{T} = (2^{-1/2}, \text{sen } t, \cos t, \text{sen } 2t, \cos 2t, \ldots)$. Assim a variância não será constante ao longo do gráfico. Os testes de significância podem continuar a ser elaborados para valores específicos de t, modificando (3.6),

$$Z = \frac{\{f_{\mathbf{x}}(t) - f_{\mu_0}(t)\}}{\sqrt{\mathbf{u}^{T}\,\mathbf{W}\,\mathbf{u}}}\,. \tag{3.9}$$

Z possui, neste caso, distribuição t-de Student com os mesmos graus de liberdade de **W**. Porém, o teste global já não é apropriado.

As estatísticas referidas não serão consideradas aqui uma vez que estão fora do espírito do presente texto que vai no sentido da

exploração dos dados.

Pode ser útil examinar os valores da função associados com um valor particular de t, por exemplo, t_0. Andrews mostra que $f_{\mathbf{x}}(t_0)$ é proporcional ao comprimento da projecção do vector $\mathbf{x}$ no vector

(3.10) $$\mathbf{u}_0^T = (2^{-1/2}, \text{sen } t_0, \cos t_0, \text{sen } 2t_0, \cos 2t_0, \ldots),$$

ou seja, $f_{\mathbf{x}}(t_0) = \mathbf{x}^T\mathbf{u}_0$. A projeccção neste espaço unidimensional revela, por vezes, padrões de comportamento que ocorrem só neste subespaço e que são obscurecidas por outras dimensões. Assim é sempre vantajoso examinar as curvas para ver se existem valores particulares de t onde os padrões de comportamento sejam evidentes.

As curvas de Andrews apresentam duas desvantagens, aliás, comuns aos métodos anteriores. Examinando a forma da função $f_{\mathbf{x}}(t)$ em (3.1) verifica-se que as variáveis não são igualmente ponderadas porque algumas delas estão associadas a componentes cíclicos que têm uma alta frequência e outras aos que têm uma baixa frequência. Assim, permutações na ordem das variáveis determinarão imagens diferentes. Nestes gráficos as baixas frequências são melhor apercebidas que as altas frequências pelo que será útil considerar X_1 a variável mais importante, X_2 a segunda mais importante e assim por diante. Na ausência de ideias firmes sobre a ordem das variáveis pode efectuar-se, previamente,

uma análise de componentes principais fazendo corresponder X_1 ao primeiro componente, X_2 ao segundo componente, etc.

Os gráficos não são operacionais quando o conjunto de dados comporta muitas observações porque o desenho se pode tornar confuso e a identificação das curvas é difícil. Neste caso, recomenda-se uma representação de todas as observações num gráfico para extrair as características gerais da amostra e, depois, fazer gráficos separados, de, por exemplo dez observações, estudando em cada um destes as curvas individuais.

A existência de variância comum para as variáveis e a não correlação entre elas são hipóteses, em geral, pouco realistas. Gnanadesikan(1977) e Embrechts e Herzberg(1991) propõem a estandardização dos dados provando, estes últimos, que as propriedades i), ii) e iii) se mantêm apesar das variáveis não perderem a correlação. Embrechts e Herzberg(1991) referem, ainda, que agrupar variáveis muito correlacionadas e colocar variáveis com grande poder discriminante nas frequências extremas, salienta as características dos dados.

As propriedades i), ii) e iii) verificam-se para quaisquer funções ortogonais e não apenas para as funções de Fourier. Embrechts e Herzberg estudaram duas funções polinomiais: o polinómio de Chebychev de grau n, $T_n(x)$ e o polinómio de Legendre de grau n, $P_n(x)$ e concluiram que estes dois tipos de

representações não acrescentavam informação relativamente às funções de Fourier.

As figuras 3.8 a 3.13 mostram as curvas de Andrews aplicadas ao Exemplo. Foram ensaiadas duas ordenações de variáveis, designadas permutações A e B que se descrevem de seguida:

Permutação A - $X_1 \equiv$ CPPNB, $X_2 \equiv$ ENRG, $X_3 \equiv$ EPVN, $X_4 \equiv$ PURB, $X_5 \equiv$ PIBA e $X_6 \equiv$ HBMED;

Permutação B - $X_1 \equiv$ HBMED, $X_2 \equiv$ PIBA, $X_3 \equiv$ PURB, $X_4 \equiv$ EPVN, $X_5 \equiv$ ENRG e $X_6 \equiv$ CPPNB.

A escolha das ordenações anteriores baseou-se na correlação existente entre as variáveis de forma que correlação elevada justificou proximidade das variáveis e correlação fraca afastamento das variáveis.

A sobreposição das 30 curvas referentes aos 30 países torna os mapas pouco elucidativos pelo que se preencherem os gráficos com dois tipos de traço: contínuo para as curvas correspondentes aos países da Europa e tracejado para as curvas correspondentes aos países de África.

Faça-se, agora, a leitura das figuras. Da figura 3.8 depreende-se a existência de dois grupos de países: precisamente os países de

FIGURA 3.8 — As curvas de Andrews para 30 paises de Africa e Europa (Permutacao A)

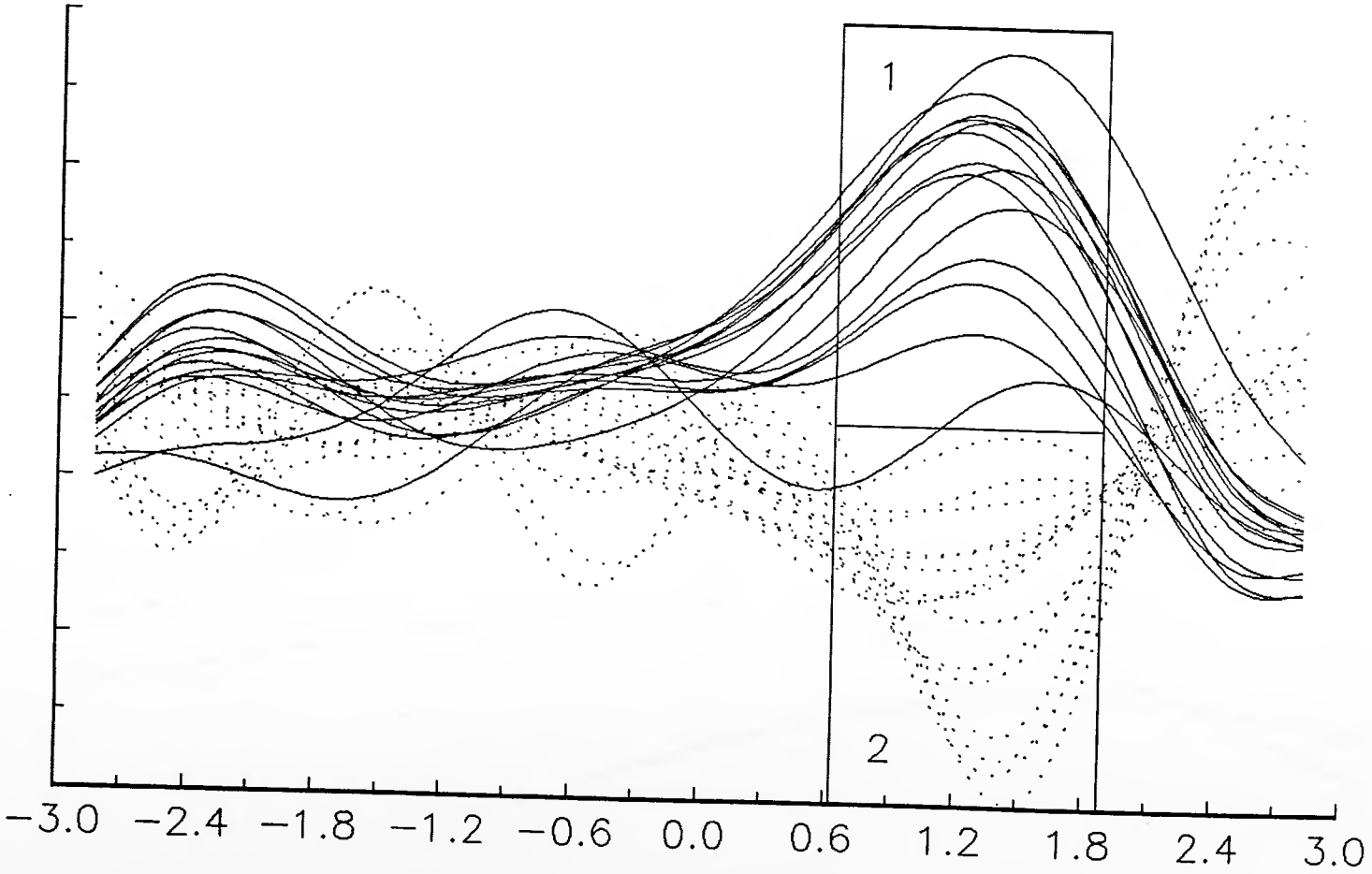


FIGURA 3.9 – As curvas de Andrews (cont.) – Sector 1 (Permutacao A)

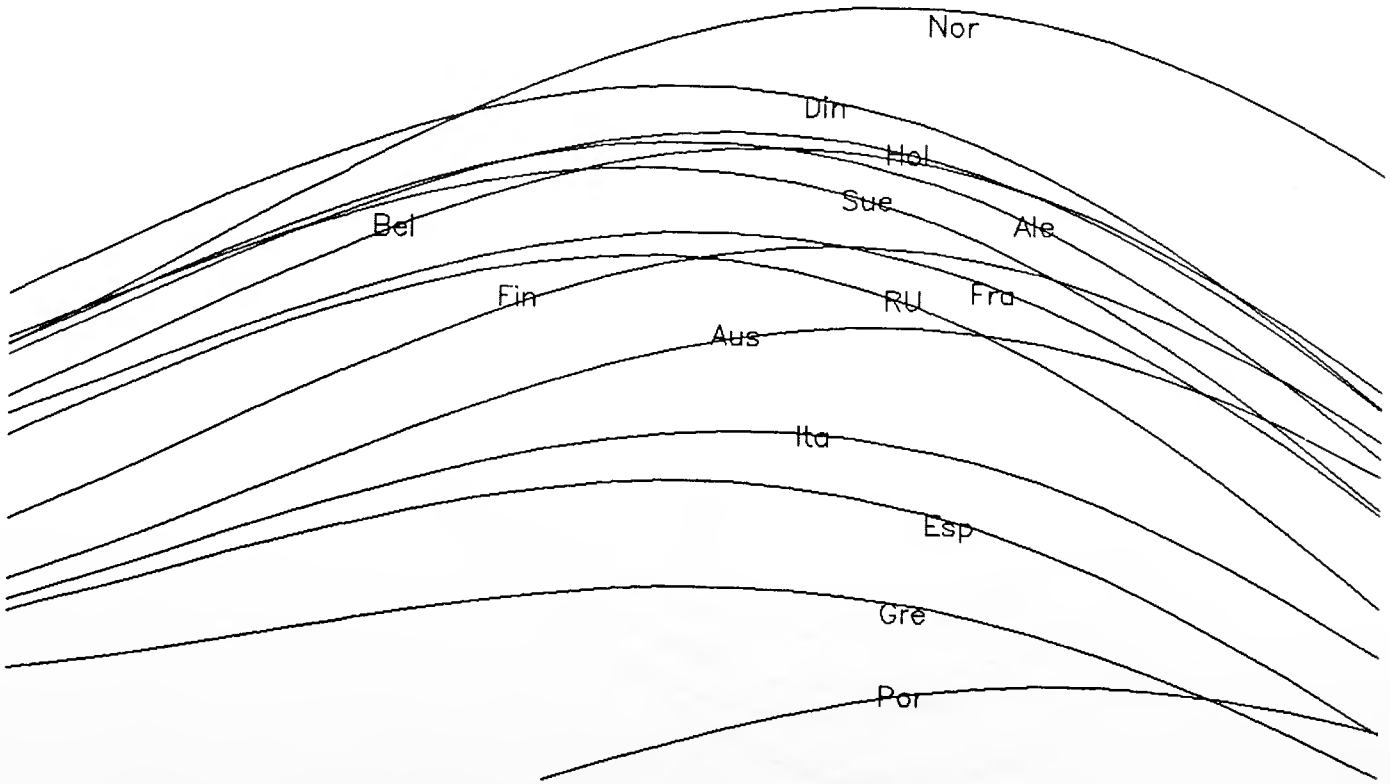


FIGURA 3.10 – As curvas de Andrews (cont.) – Sector 2 (Permutacao A)

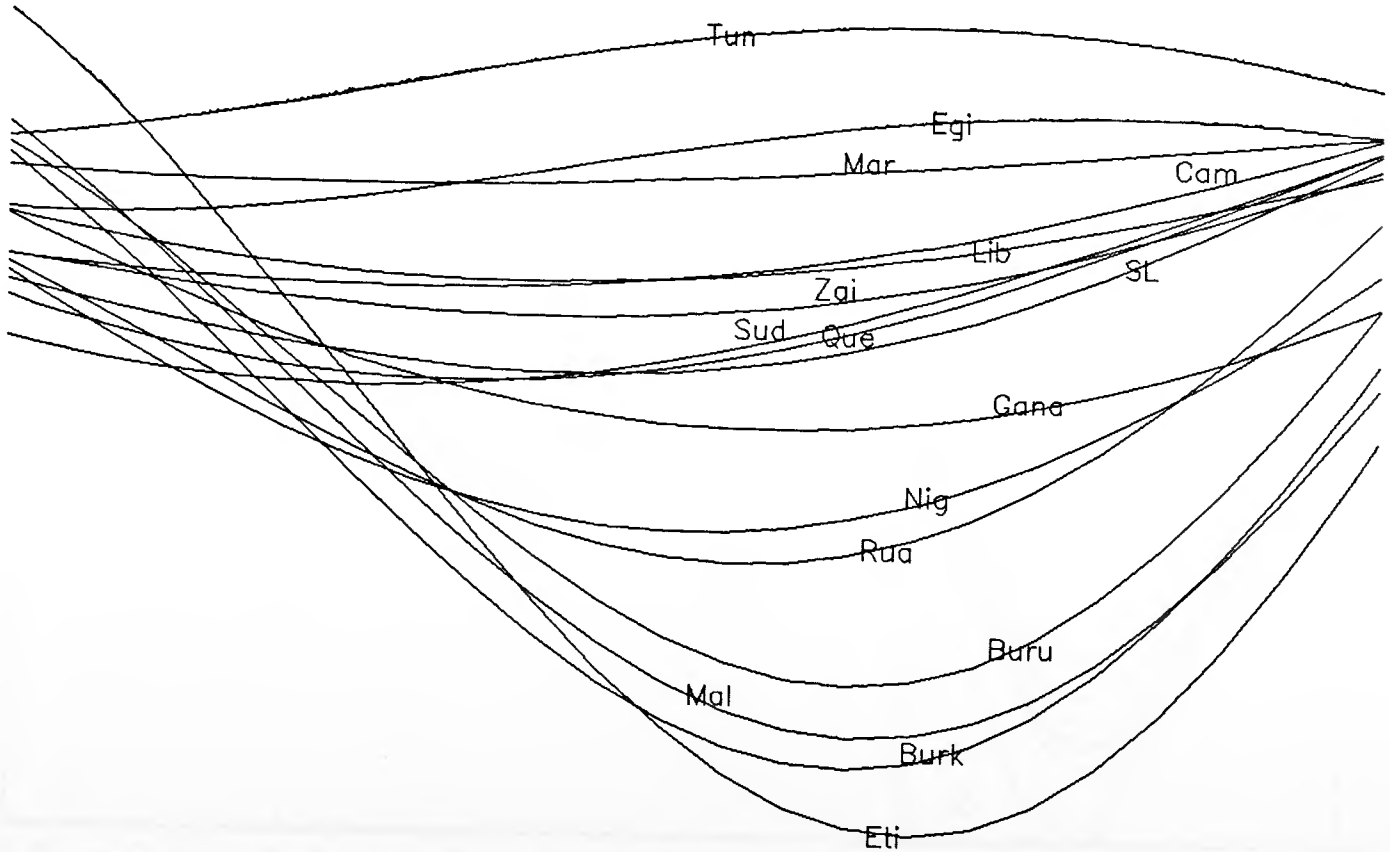


FIGURA 3.11 – As curvas de Andrews para 30 paises de Africa e Europa
(Permutacao B)

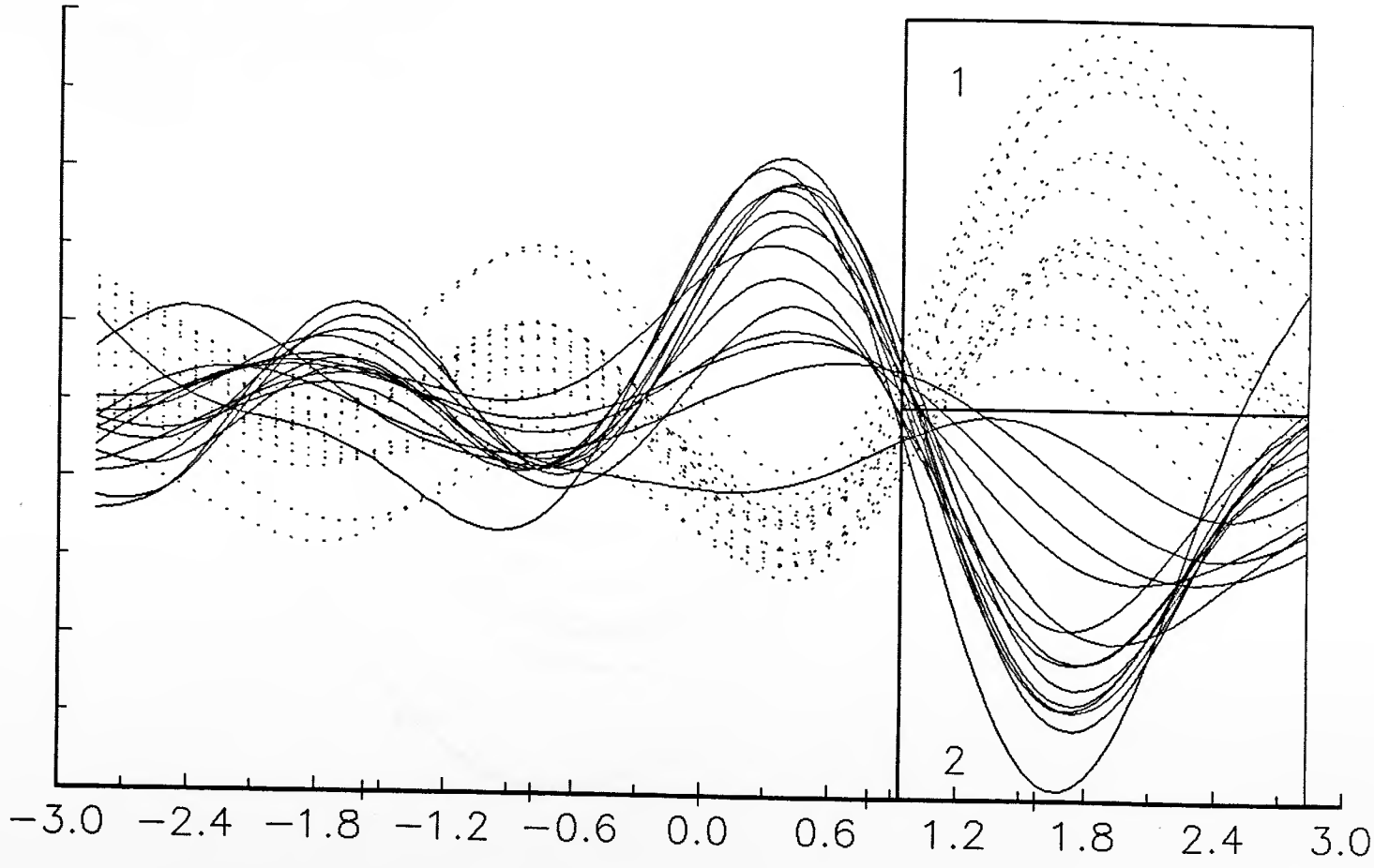


FIGURA 3.12 – As curvas de Andrews (cont.) – Sector 2 (Permutacao B)

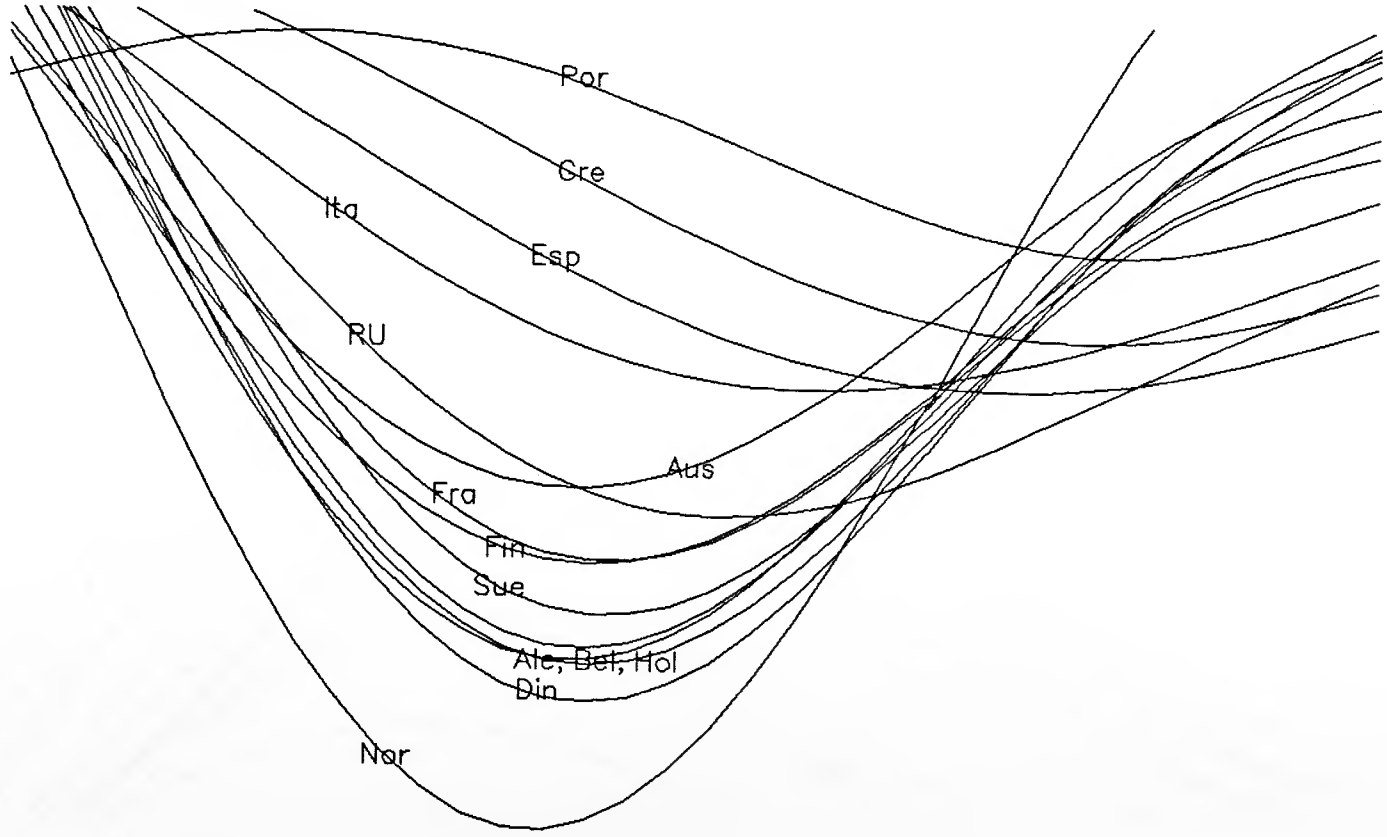
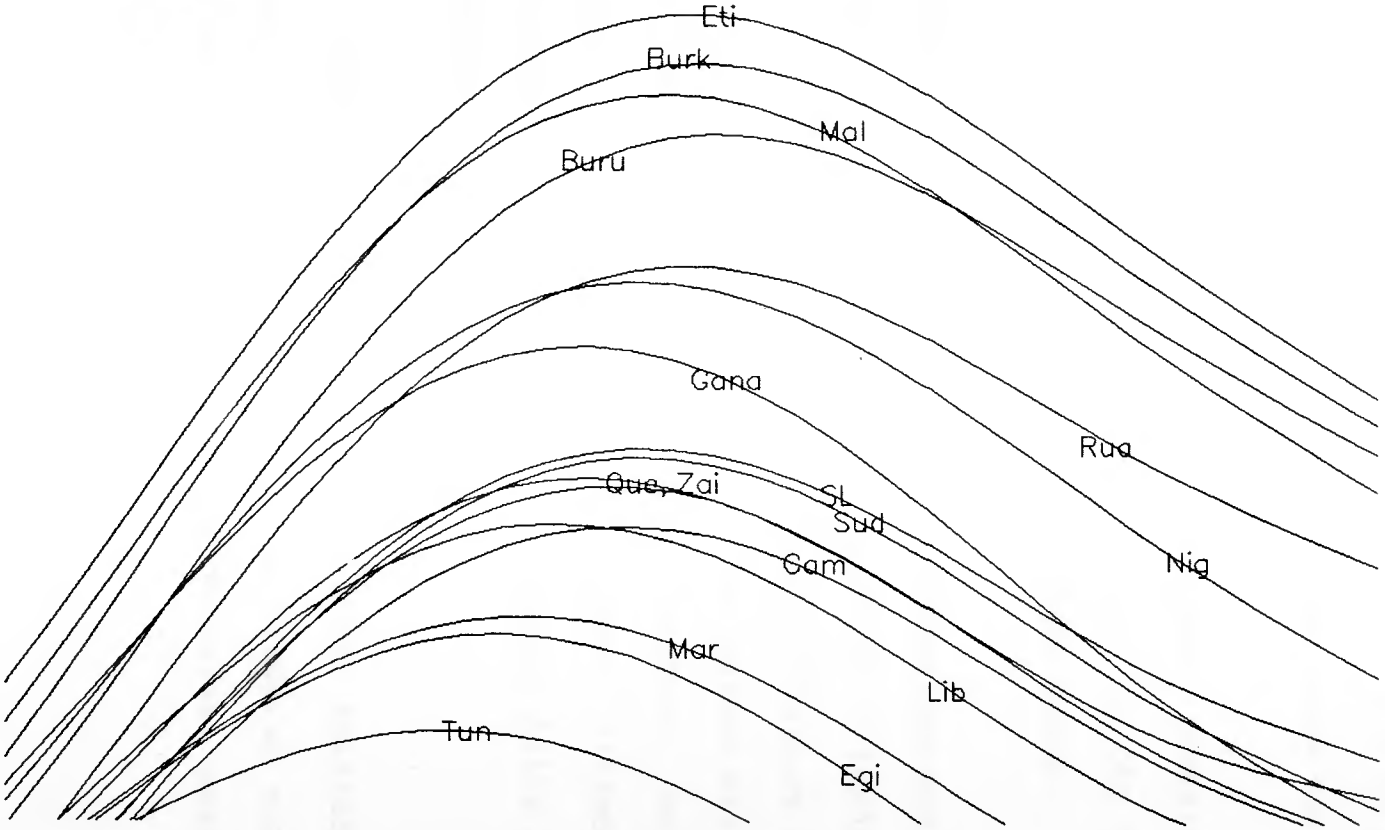


FIGURA 3.13 – As curvas de Andrews (cont.) – Sector 1
(Permutacao B)



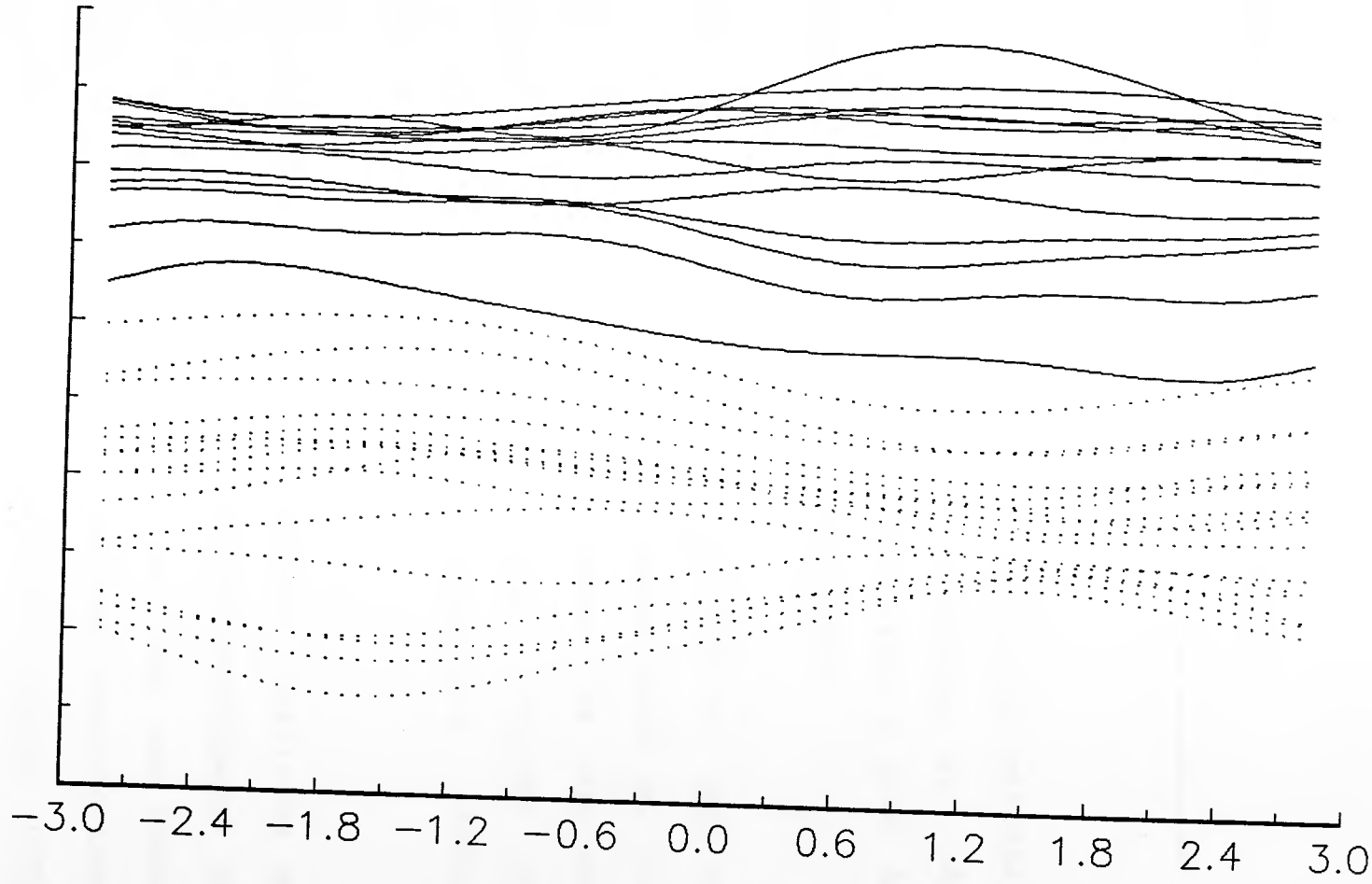
África e da Europa. Cada um dos grupos tem comportamento idêntico para todos os valores de t e os dois grupos distanciam-se para valores de t entre, aproximadamente 0.6 e 1.8. Os sectores sublinhados constituem os gráficos das figuras 3.9 e 3.10 onde é feita a identificação dos países representados por cada curva.

A figura 3.11 ilustra a permutação B e, de novo, ressaltam os dois grupos: Europa e África. Nos gráficos das figuras 3.12 e 3.13 são identificados os países associados a cada curva.

As conclusões retiradas do conjunto dos mapas referidos confirmam as ideias expostas ao longo do presente capítulo oriundas das técnicas simbólicas e que salientam a separação entre os países da Europa do Norte e Europa do Sul e no seio dos países africanos o isolamento do grupo dos infelizes Etiópia, Burkina, Malawi e Burundi. O Ruanda, acompanhante habitual destes últimos, é, segundo o método das curvas de Andrews um pouco mais feliz porque se afasta ligeiramente do grupo.

A figura 3.14 resulta de previamente ter sido ensaiada uma análise de componentes principais. As curvas perderam algum do aspecto cíclico mas continuam a distinguir os países segundo o continente em que estão localizados.

FIGURA 3.14 – As curvas de Andrews para 30 paises de Africa e Europa
(Com Analise de Componentes Principais)



CAPÍTULO 4

REPRESENTAÇÃO DE MATRIZES DE PROXIMIDADE

A matéria-prima do presente capítulo é constituída pelas matrizes de proximidades que, como se referiu no capítulo **2.**, são caracterizadas por descreverem relações, reflectindo o grau de semelhança ou dissemelhança, entre pares de entidades.

A análise traçada dirige-se, sobretudo, ao caso em que as entidades são os indivíduos mas não perde generalidade se for feita para matrizes de associação entre variáveis. Por isso não se faz, ao longo do texto, distinção entre matrizes de proximidades Δ_I e matrizes de associação Δ_V(vidé cap.**2.**), sendo utilizada a notação genérica de Δ.

As matrizes de proximidades podem ser calculadas a partir de um conjunto de dados multivariados ou resultarem directamente da recolha de dados. Tradicionalmente, neste último caso, são dados subjectivos obtidos através de inquéritos a pessoas, pedindo-lhes que julguem a dissemelhança entre pares de coisas como automóveis, candidatos políticos, tipos de vinho, etc.. Mas podem

ser também medidas objectivas como tempo de condução entre pares de cidades ou a frequência com que pares de pessoas de um grupo falam entre elas, etc..

A análise de conglomerados usa matrizes de proximidade para dividir os indivíduos num pequeno número de grupos relativamente homogéneos enquanto que o "*Multidimensional Scaling*" (MDS), de que a seguir se dá conta, permite examinar as posições relativas de todos os sujeitos ao representar a estrutura dos dados como uma figura geométrica.

O termo MDS refere-se a uma família de métodos através dos quais a informação contida num conjunto de dados é representado por um conjunto de pontos num espaço. Estes pontos estão dispostos de tal forma que as relações geométricas, tais como distância entre pontos, reflectem relações de semelhança ou dissemelhança entre os dados. Assim, objectos semelhantes são representados por pontos próximos um do outro e objectos dissemelhantes por pontos afastados. Um bom modelo é aquele em que a distância entre dois pontos seja igual, ou o mais aproximada possível, à dissemelhança entre as duas entidades por eles representadas. Estabelecem-se, deste modo, correspondências cruciais entre entidades e pontos e entre dissemelhanças e distâncias.

As técnicas de MDS desenvolveram-se sobretudo no âmbito das ciências sociais e em particular no da psicologia quantitativa.

Eis alguns exemplos de áreas onde a aplicação de MDS é frequente: os psicólogos tentam comprender a percepção e avaliação de estímulos auditivos (voz ou sons musicais), visuais (cores ou faces) ou entidades sociais (traços de personalidade ou situações sociais); os sociólogos determinam a estrutura de grupos e organizações segundo as percepções dos membros; os antropólogos comparam grupos culturais diferentes, baseados nas suas crenças ou linguagem; os economistas investigam as reacções de consumidores a uma grande variedade de classes de produtos; os pedagogos estudam a estrutura da inteligência e dos ambientes escolares, etc..

A história do MDS divide-se em três etapas, correspondendo cada uma ao desenvolvimento de um método de análise inovador. Os três tipos assim definidos são: MDS clássico ou métrico, MDS ordinal ou não métrico e MDS de diferenças individuais.

O objectivo de qualquer técnica de MDS é produzir uma figura geométrica que saliente a estrutura dos dados e se caracterize pela simplicidade e facilidade de compreensão. As diferenças entre os métodos registam-se ao nível da forma como as proximidades estão medidas, do número de matrizes englobadas e do modelo geométrico considerado.

O MDS clássico (MDSC) constrói a configuração dos pontos a partir de uma só matriz de proximidades que se encontram medidas numa

escala numérica (por isso o MDS é métrico) e cuja configuração tem as propriedades de um modelo geométrico euclideano. A ideia básica foi introduzida por Richardson(1938) embora o primeiro procedimento métrico e a denominação "*Multidimensional Scaling*" se devam a Torgerson(1958). Gower(1966) popularizou a técnica sob o nome de análise de coordenadas principais.

O MDS ordinal (MDSO) distingue-se do método anterior porque usa apenas a ordenação das proximidades. Esta técnica foi desenvolvida por Shepard(1962a,b) e Kruskal(1964a,b) que, além da produção teórica, elaboraram um algoritmo para encontrar a representação de pontos desejada. Diversas alternativas a este algoritmo têm sido estudadas desde então, mas todas respeitam, no essencial, o que foi proposto por Kruskal.

O MDS de diferenças individuais(MDSD), descrito no capítulo 6, é apropriado para analisar, simultaneamente, mais que uma matriz de proximidades podendo, estas, estar medidas numa escala ordinal ou numérica. Diferentes correntes de MDSD configuram o espaço resultante da análise segundo um modelo euclideano ou euclideano ponderado.

Nos últimos anos o MDS extendeu-se na vertente de MDSD pelo que Young e Harris(1990) apelidaram de clássicos os dois métodos que usam só uma matriz, MDSC e MDSO, distinguindo-os, unicamente, pela escala de medida das proximidades.

4.1. "*MULTIDIMENSIONAL SCALING*" CLÁSSICO

O MDSC segue um procedimento a dois passos, construindo algebricamente, no primeiro, uma configuração de pontos com base nas dissemelhanças entre os objectos. Mas a configuração pode ter demasiadas dimensões para que seja compreensível, pelo que se torna necessário fazer, no segundo passo, a identificação de um subespaço no qual os pontos possam ser projectados e onde as dissemelhanças sejam aproximadas o melhor possível por distâncias correspondentes.

Se for dado um conjunto de n pontos num espaço euclideano p-dimensional, P_i (i=1,...n), cujas coordenadas x_{ik} (i=1,...,n; k=1,...p) são os elementos da matriz de dados original, pode, facilmente, calcular-se as distâncias euclideanas quadráticas d^2_{ij} (i,j=1,...n), entre quaisquer pares de pontos P_i e P_j, usando a fórmula habitual,

$$d^2_{ij} = \sum_{k=1}^{p} (x_{ik}-x_{jk})^2, \quad (i,j=1,\ldots,n). \tag{4.1}$$

O problema que o MDSC visa responder põe-se do seguinte modo: se, inversamente, se conhecem as distâncias quadráticas de um conjunto de n pontos nalgum espaço euclideano, como obter as

coordenadas x_{ik}? Este método é, por assim dizer, o inverso do processo de cálculo de Δ a partir de $\mathbf{X}$, no caso em que $\delta_{ij} = d_{ij}^2$.

Seja, então, Δ a matriz que contém os elementos d_{ij}^2. Os valores de x_{ik} são obtidos a partir de d_{ij}^2 construindo uma matriz $\mathbf{B}$ definida por,

$$\mathbf{B} = \mathbf{X}\mathbf{X}^T. \tag{4.2}$$

O elemento genérico de $\mathbf{B}$ é,

$$b_{ij} = \sum_{k=1}^{p} x_{ik}x_{jk}, \quad (i,j=1,\ldots,n), \tag{4.3}$$

pelo que as distâncias d_{ij}^2 podem ser escritas como,

$$d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}. \tag{4.4}^*$$

O objectivo, neste momento, é inverter esta última equação de modo a obter os valores b_{ij} a partir de d_{ij}^2. Não existe, no entanto, uma solução única: obtida uma configuração de pontos com as distâncias requeridas é possível calcular um outro conjunto de coordenadas, igualmente válidas, ao dar outra disposição aos pontos através de translações ou rotações da configuração. Ao executar estas operações, as distâncias euclideanas não são

alteradas(ver cap.7.). Chatfield(1980) descreve, empiricamente, a situação:

> *"...Se se conhecessem as distâncias entre todas as cidades inglesas, mesmo assim nada se sabia àcerca da latitude ou longitude de Londres nem se Penzance se situa no norte, no sul, este ou oeste mas apenas que se trata de um ponto extremo."*

Por outras palavras, não se consegue determinar, unicamente, nem a localização nem a orientação da figura. É, pois, conveniente remover esta indeterminação impondo uma restrição: obriga-se o centróide do conjunto dos n pontos a situar-se na origem das coordenadas, o que equivale a impor,

$$(4.5)^* \qquad \sum_{i=1}^{n} x_{ik} = 0 \ , \qquad (k=1,\ldots,p).$$

Embora o problema de localização seja resolvido ao colocar o centro de gravidade da configuração na origem, o problema de orientação não fica resolvido porquanto a configuração pode, ainda, ser sujeita a qualquer transformação ortogonal e as distâncias permanecerão inalteradas.

Com a restrição (4.5), os valores b_{ij} obtêm-se como,

(4.6)* $$b_{ij} = -\frac{1}{2}\left[d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2 \right] ,$$

em que, na notação habitual, $d_{i.}^2$ é a média da i-ésima linha, $d_{.j}^2$ a média da j-ésima coluna e $d_{..}^2$ a média de todos os elementos de Δ.

Concluída a construção da matriz **B** passa-se à determinação da matriz de coordenadas **X**. Através de (4.2) verifica-se que **B** é matriz simétrica podendo ser expressa de acordo com a conhecida decomposição espectral de uma matriz,

(4.7)* $$\mathbf{B} = \mathbf{VLV}^T,$$

onde **V** é matriz ortogonal cujas colunas são os vectores próprios de **B**, $\mathbf{v}_i$ (i=1,..,n), correspondentes aos valores próprios de **B**, λ_i. Os valores próprios constituem a matriz diagonal **L** e apresentam-se dispostos por ordem decrescente de magnitude ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$). Os vectores $\mathbf{v}_i$ estão normalizados e são ortogonais entre si ($\mathbf{v}_i^T\mathbf{v}_i$=1 e $\mathbf{v}_i^T\mathbf{v}_j$=0 para $i \neq j$ e i,j=1,...,n). Devido à condição de ortogonalidade os valores próprios são todos distintos.

Se **B** é matriz definida positiva tem valores próprios λ_i positivos e as raízes quadradas dos λ_i são reais. Igualando (4.2) e (4.7) obtém-se a matriz de coordenadas **X**,

$$(4.8)^{*} \qquad \mathbf{X} = [\sqrt{\lambda_1}\,\mathbf{v}_1 \quad \sqrt{\lambda_2}\,\mathbf{v}_2 \quad \ldots \quad \sqrt{\lambda_n}\,\mathbf{v}_n] \; .$$

Se **B** é semidefinida positiva de característica k ($k \leq n$), então **B** tem k valores próprios positivos e $(n-k)$ nulos, e a matriz **X** é k-dimensional,

$$\mathbf{X} = [\sqrt{\lambda_1}\,\mathbf{v}_1 \quad \sqrt{\lambda_2}\,\mathbf{v}_2 \quad \ldots \quad \sqrt{\lambda_k}\,\mathbf{v}_k] \; .$$

Se **B** tem característica n e se deseja obter uma solução numa dimensão t, ($t<n$), tão pequena quanto possível, deve considerar-se a matriz **L** e construir-se **X** a partir dos primeiros t valores próprios, que são, ao mesmo tempo, os maiores. Esta solução, proposta por Eckart e Young(1936) (vidé Anexo-(4.7)), é a que minimiza a soma dos quadrados da diferença entre as coordenadas das configurações n e t-dimensionais.

O exemplo seguinte retirado de Gordon(1981) ilustra o MDSC. Seja a matriz Δ que contém as distâncias euclideanas quadráticas d_{ij}^2 entre cinco pontos, P_i (i=1,...,5):

$$\Delta = \begin{bmatrix} 0 & 4 & 5 & 16 & 20 \\ & 0 & 5 & 20 & 16 \\ & & 0 & 5 & 5 \\ & & & 0 & 4 \\ & & & & 0 \end{bmatrix} \; .$$

A matriz **B** calculada por (4.6) é

$$\mathbf{B} = \begin{bmatrix} 5 & 3 & 0 & -3 & -5 \\ 3 & 5 & 0 & -5 & -3 \\ 0 & 0 & 0 & 0 & 0 \\ -3 & -5 & 0 & 5 & 3 \\ -5 & -3 & 0 & 3 & 5 \end{bmatrix}$$

Esta matriz possui dois valores próprios positivos e três nulos. Os valores próprios positivos e os vectores próprios normalizados correspondentes são:

$$\lambda_1=16, \qquad \mathbf{v}_1 = \begin{bmatrix} -1/2 \\ -1/2 \\ 0 \\ 1/2 \\ 1/2 \end{bmatrix},$$

e,

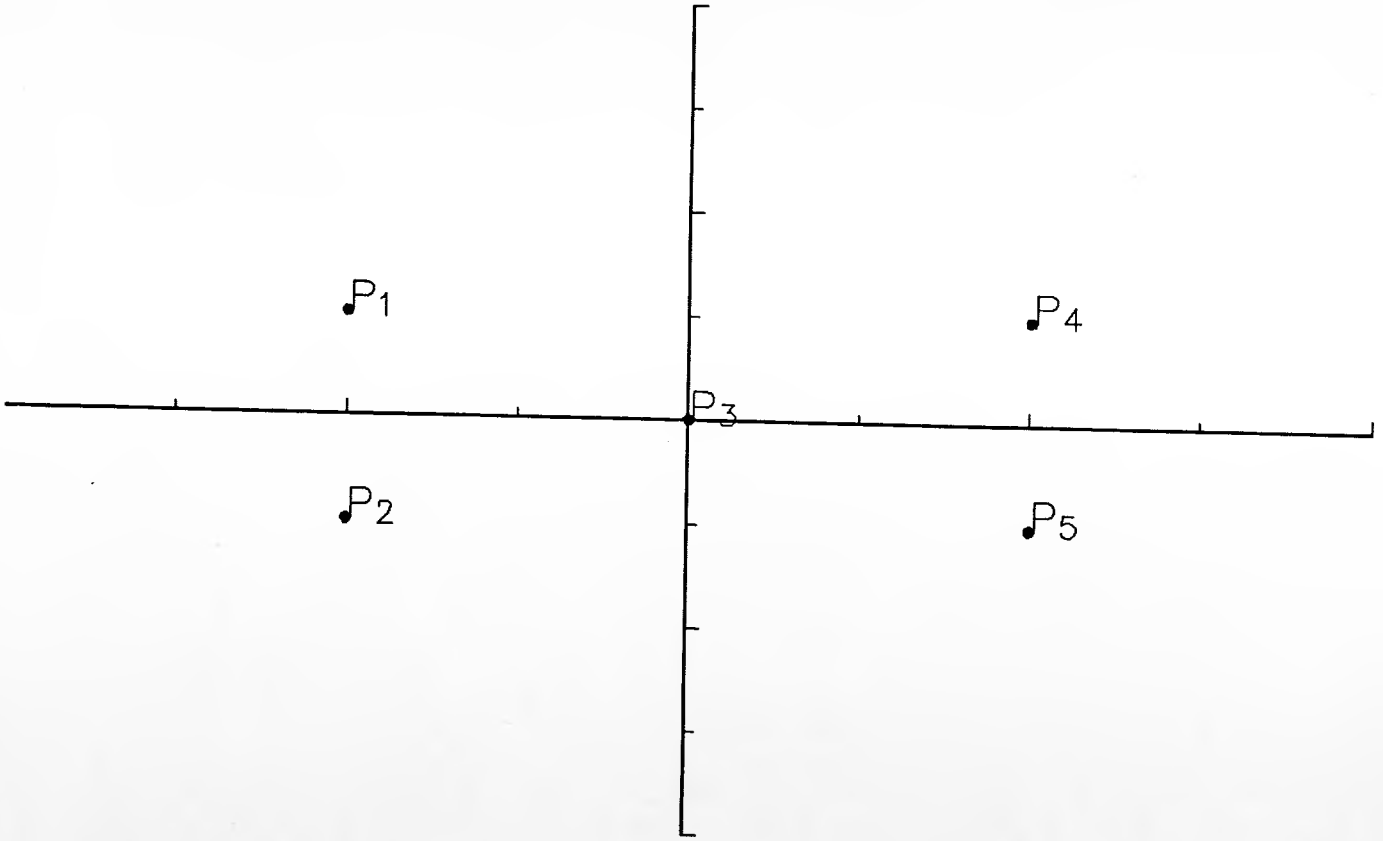
$$\lambda_2=4, \qquad \mathbf{v}_2 = \begin{bmatrix} 1/2 \\ -1/2 \\ 0 \\ 1/2 \\ -1/2 \end{bmatrix}.$$

Assim,

$$\mathbf{X} = \begin{bmatrix} -2 & 1 \\ -2 & -1 \\ 0 & 0 \\ 2 & 1 \\ 2 & -1 \end{bmatrix} \begin{matrix} (P_1) \\ (P_2) \\ (P_3) \\ (P_4) \\ (P_5) \end{matrix}$$

As coordenadas do *i*-ésimo ponto são dadas pelos valores da *i*-ésima linha da matriz **X**. A configuração bidimensional destes cinco pontos (figura 4.1) reproduz as distâncias interpontos situando o centróide dos pontos na origem das coordenadas. A

FIGURA 4.1 – Solucao de MDSC para o exemplo artificial



leitura do gráfico faz-se interpretando as distâncias relativas entre os vários pontos. P_3 está igualmente distanciado dos outros quatro pontos significando que a entidade por ele representada é igualmente dissemelhante das outras quatro entidades. P_1 está mais próximo de P_2 do que de P_4, portanto, a entidade representada por P_1 tem características que se assemelham mais à entidade representada por P_2 do que à de P_4.

Como já foi referido, é impossível determinar uma só orientação para a representação das distâncias. Com efeito, a orientação pode ser modificada por uma transformação ortogonal arbitrária de modo a encontrar novos eixos de referência.

O facto de as coordenadas serem obtidas a partir dos valores próprios e vectores próprios da matriz **B** significa que a representação se refere aos eixos principais. Tal como sucede na análise de componentes principais(ACP), também as melhores *t* dimensões nas quais se representa a amostra são dadas pelos vectores próprios correspondendo aos primeiros *t* valores próprios. Uma vez que os valores próprios foram dispostos por ordem decrescente, os sucessivos vectores próprios referem-se a dimensões de importância decrescente. A ACP parte da matriz de dados originais **X** e calcula os valores e vectores próprios da matriz de covariâncias da amostra que é proporcional a $\mathbf{X}^T\mathbf{X}$. Ora, os valores próprios de $\mathbf{X}^T\mathbf{X}$ são os mesmos de $\mathbf{X}\mathbf{X}^T$, isto é, de **B**, enquanto que entre os vectores próprios das duas matrizes-produto

existe uma relação linear. Sejam $\mathbf{v}_j$ e $\mathbf{u}_i$, respectivamente, os vectores próprios de $\mathbf{X}^T\mathbf{X}$ e de $\mathbf{X}\mathbf{X}^T$ que se relacionam,

$$(4.9)^* \qquad \mathbf{v}_j = \mathbf{X}\,\mathbf{u}_i \qquad (i=1,\ldots,n;j=1,\ldots,p),$$

embora sejam de dimensões diferentes: $\mathbf{v}_j$ são vectores (px1) enquanto $\mathbf{u}_i$ são vectores (nx1).

Assim os resultados duma ACP são exactamente equivalentes aos do MDSC se as dissemelhanças calculadas a partir da matriz de dados forem distâncias euclideanas. A equivalência entre as duas técnicas significa que, se se pretender reduzir a dimensão de uma matriz de dados através de uma transformação em novos eixos não é necessário efectuar ambas as análises. No caso de $n>p$ é preferível uma ACP porque é mais fácil encontrar os vectores próprios da matriz $\mathbf{X}^T\mathbf{X}$ (pxp) do que da matriz $\mathbf{X}\mathbf{X}^T$(nxn). Em vez de equivalência entre os dois métodos deve, antes, falar-se de dualidade: a ACP é uma técnica-*R* enquanto a análise de coordenadas principais é uma técnica-*Q*.(vidé cap.**2.**). O MDSC torna-se a técnica apropriada quando não é conhecida a matriz de dados original mas apenas a matriz de proximidades.

Uma grande vantagem deste método é poder ser usado para encontrar um conjunto de coordenadas para as observações quando as dissemelhanças não são euclideanas. Quando **B** é calculada a partir de distâncias euclideanas, **B** é, seguramente, semidefinida

positiva pelo que tem valores próprios não negativos e as coordenadas (4.8) são reais. Se a matriz **B** não for calculada a partir de distâncias euclideanas não é, necessariamente, semidefinida positiva pelo que, alguns dos n valores próprios encontrados podem ser negativos. Há, no entanto, pelo menos um valor próprio nulo uma vez que a soma de elementos em cada linha de **B** é zero. A existência de valores próprios negativos conduz a configurações de dimensões imaginárias (por (4.8)). Mardia(1978) conclui que, se os valores próprios negativos forem poucos e de pequena magnitude, a representação é possível dando atenção só aos valores próprios positivos e respectivos vectores próprios. Sibson(1979) determinou que MDSC é, surpreendemente, robusto a distorções das distâncias euclideanas. Porém, se a matriz **B** tem, pelo menos, um valor próprio negativo grande, o MDSC não deve ser aplicado porque a solução tende a dispersar-se num grande número de dimensões. Neste caso, as distâncias não são, nem aproximadamente, euclideanas podendo, em alternativa, usar-se MDS ordinal.

Como se referiu, se se procurar uma configuração num determinado número de dimensões t, consideram-se apenas os t valores próprios maiores e os vectores correspondentes, calculando-se a matriz de coordenadas com a equação (4.8). A medida da distorção causada pela redução do número de dimensões de n para t é dada por,

$$(4.10)^{*} \qquad G_1 = \left(\sum_{i=1}^{t} \lambda_i \right) / \left(\sum_{i=1}^{n} \lambda_i \right).$$

A medida G_1 assume valores no intervalo real [0,1], indicando os valores próximos de 1 que se está em presença de um bom ajustamento. No caso da matriz **B** possuir valores próprios negativos G_1 não deve ser usada porque o denominador é menor que a soma dos valores próprios positivos, o que dá uma visão demasiado optimista do ajustamento da configuração escolhida. Pode, então, substituir-se G_1 por,

$$(4.11) \qquad G_2 = \left(\sum_{i=1}^{t} \lambda_i \right) / \sum_{i=1}^{n} |\lambda_i| \; .$$

A "*Minimum Spanning Tree*" (MST) da matriz Δ é um útil instrumento gráfico para avaliar se as distâncias originais são preservadas pela configuração a duas dimensões.

A MST, socorrendo-se de algumas ideias básicas da teoria dos grafos, associa cada sujeito a um ponto ou vértice num espaço e associa a dissemelhança δ_{ij} entre pares de sujeitos à distância, ou comprimento do segmento, entre os correspondentes pares de pontos.

Dados *n* pontos em várias dimensões a MST destes pontos é qualquer

conjunto de segmentos de recta unindo pares de pontos que representam entidades semelhantes e satisfazem as condições:

1. Não ocorrem "*loops*" fechados.
2. Cada ponto é visitado pelo menos por uma linha.
3. A árvore está toda ligada, ou seja, todos os pontos estão ligados entre si, directa ou indirectamente.
4. Sendo o comprimento de uma árvore definido como a soma do comprimento dos seus segmentos, a MST tem o menor comprimento.

Gower e Ross(1969) discutem vários algoritmos para encontrar a MST e o mais popular opera iterativamente. Constituem-se dois conjuntos; o conjunto A contém os segmentos que pertencem à MST e o conjunto B os que não pertencem. De início A está vazio e a operação consiste em escolher o menor segmento de B. Na segunda e restantes iterações escolhe-se o menor segmento de B que, ao mesmo tempo, não forme "*loop*" fechado com qualquer dos segmentos de A. O algoritmo encerra-se no momento em que A contenha n-1 segmentos.

Sobrepõem-se os ramos da MST de Δ no gráfico a duas dimensões que resultou da aplicação de uma técnica de ordenação. Os pontos que se encontram próximos na configuração mas não estão ligados pela MST, constituem distorções.

Se bem que as denominações "MDSC", "Análise de coordenadas

Principais" e "MDS métrico" sejam, na realidade, sinónimos quando foram propostos continham ligeiras diferenças. Gower(1966), ao elaborar uma ARP, parte de uma matriz de semelhanças em vez de uma matriz de dissemelhanças fazendo uma análise perfeitamente similar à que aqui se apresentou desde que se utilize a transformação (2.2).

O termo 'métrico' foi introduzido por Kruskal e Wish(1978) para situações em que existe alguma função analítica que relaciona as dissemelhanças e as distâncias euclideanas. No passado era prática comum pensar as proximidades como distâncias euclideanas fora a adição de uma constante, questão conhecida como 'o problema da constante-aditiva', ou, mais genericamente, as proximidades eram supostas derivar de alguma transformação linear das distâncias euclideanas,

(4.12) $$\delta_{ij} \approx d_{ij} + a \ .$$

Em resumo, o algoritmo utilizado pelo MDSC pressupõe a existência de uma matriz de proximidades, apenas satisfazendo as condições de distância(vidé **2.**) e procede-se como se fossem distâncias euclideanas:

1) Cálculo de **A** com,

$$a_{ij} = -\frac{1}{2}\delta_{ij}^2 \quad (i,j=1,\ldots,n),$$

em que δ_{ij} é o valor da dissemelhança entre os sujeitos *i* e *j*.

2) Cálculo de **B** com,

$$b_{ij} = [\, a_{ij} - a_{i.} - a_{.j} + a_{..} \,] \ ,$$

em que $a_{i.}$, $a_{.j}$ e $a_{..}$, são, respectivamente a média da *i*-ésima linha, a média da *j*-ésima coluna e a média de todos os elementos de **A**. Esta operação é designada pelos psicometristas como dupla centragem.

3) Efectuar a decomposição espectral de **B**. Sejam λ_i e $\mathbf{v}_i$ (i=1,...,n), respectivamente, os valores e os vectores próprios associados. A matriz de coordenadas é

$$\mathbf{X} = [\, \sqrt{\lambda_1}\,\mathbf{v}_1 \ \ldots \ \sqrt{\lambda_t}\,\mathbf{v}_t \,] \ .$$

Uma vez que se deseja visualizar a representação determina-se, por razões práticas, que a dimensão *t* seja igual a dois.

O MDSC é, ainda, útil ao dar uma configuração inicial para o processo iterativo usado no "*scaling*" ordinal.

A partir dos dados económico-sociais do Exemplo formou-se a matriz de dissemelhanças utilizando a medida de distância euclideana. Submeteu-se esta matriz a uma análise de MDSC cujo resultado está presente na figura 4.2 enquanto os sectores 1 e 2 assinalados são ampliados nas figuras 4.3 e 4.4. O ajustamento do gráfico às dissemelhanças é muito bom (G_1=0,99) salientando-se, da leitura do mapa, as posições extremas ocupadas pela Etiópia e pela Noruega e a proximidade de Portugal à Tunísia e ao Egipto. Curiosamente, a Etiópia encontra-se mais próximo da Noruega que do Egipto mas no caso da solução a duas dimensões ter a forma de U a estrutura dos dados é, segundo Kendall(1975), quase unidimensional. O "fenómeno da ferradura", como é conhecido, significa que a dimensão curvilínea dá uma razoável compreensão da configuração: a distância entre dois objectos é medida com precisão quando eles estão próximos mas não o é quando estão afastados porque as grandes distâncias atraem objectos distantes.

Ainda com os dados do Exemplo formou-se uma segunda matriz de dissemelhanças usando a distância "*city block*". Relativamente à distância euclideana esta medida atribui menor importância a variáveis com grande amplitude. Feita a análise de MDSC com um ajustamento óptimo (G_1=1), resultado integral visualiza-se na figura 4.5 e o sector sublinhado na figura 4.6. A Etiópia e a Noruega continuam a ocupar posições extremas mas o facto estranho deste gráfico é o isolamento da Alemanha e Suécia num local longínquo do quarto quadrante.

FIGURA 4.2 – MDSC para 30 paises de Africa e Europa
(δ_{ij}=distancia euclideana)

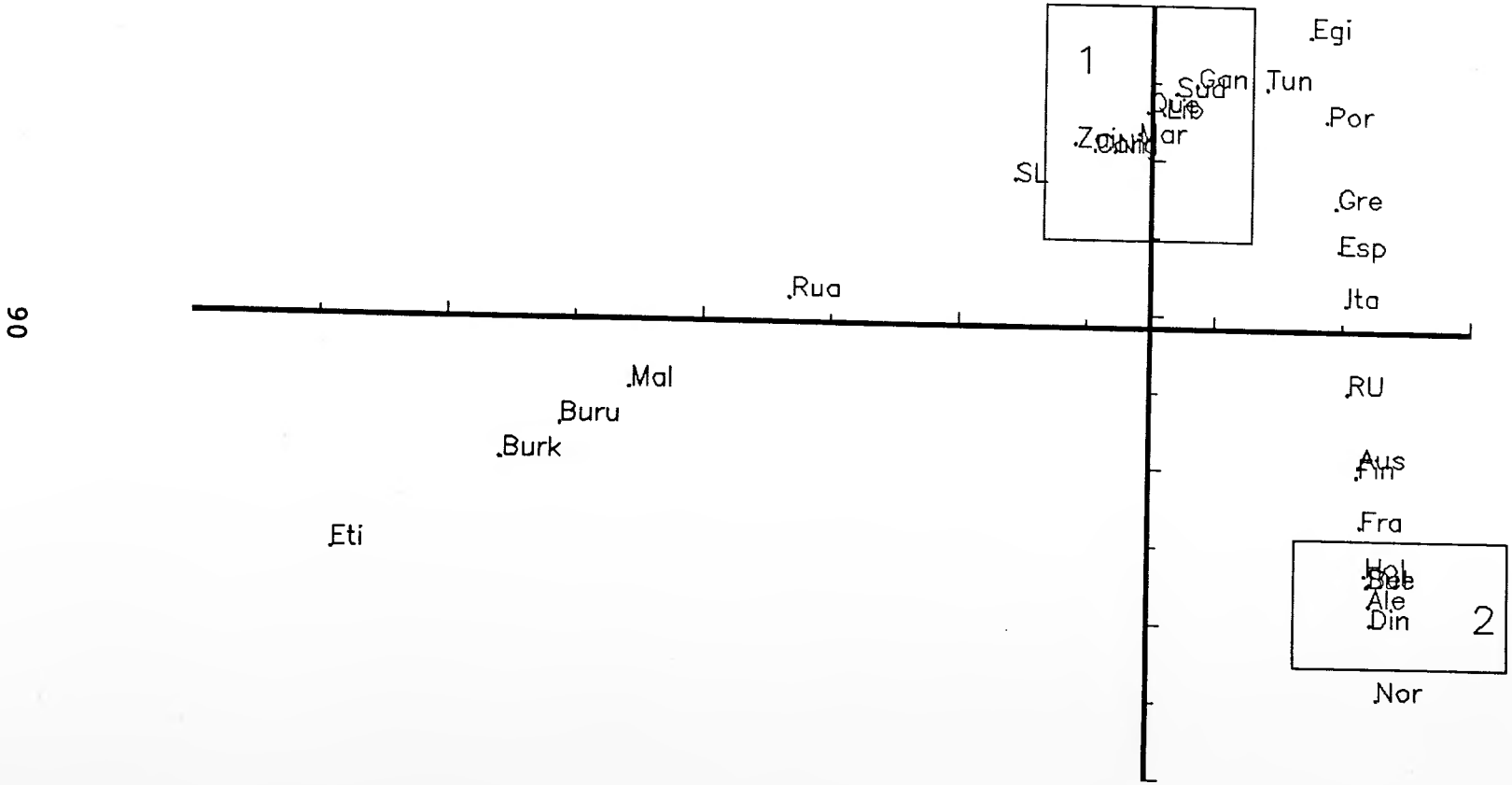


FIGURA 4.3 – MDSC para 30 paises de Africa e Europa
(δ_{ij}=distancia euclideana) – Sector 1

FIGURA 4.4 – MDSC para 30 paises de Africa e Europa
(δ_{ij}=distancia euclideana) – Sector 2

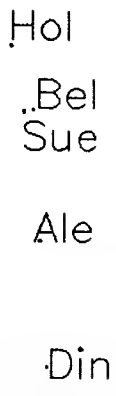
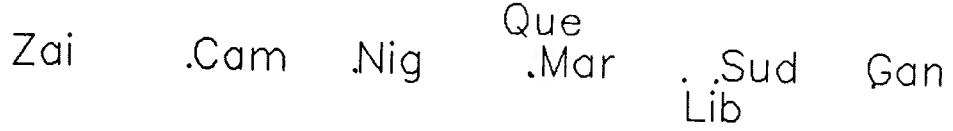


FIGURA 4.5 – MDSC para 30 paises de Africa e Europa
(δ_{ij}=distancia "*city block*")

FIGURA 4.6 – MDSC para 30 paises de Africa e Europa (δ_{ij}=distancia "*city block*") – Sector 1



Ensaiaram-se aplicações de MDSC com as medidas de dissemelhança anteriores mas ponderadas pela amplitude de cada variável e os resultados constam das figuras 4.7 e 4.10 usando, respectivamente a distância euclideana ponderada (G_2=0,93) e a distância "*city block*" ponderada (G_2=0,73). As figuras 4.8, 4.9, 4.11 e 4.12 constituem ampliações dos sectores assinalados em 4.7 e 4.10.

Os ajustamentos efectuados em qualquer das análises de MDSC anteriores são de boa qualidade pelo que se as distâncias escolhidas representarem bem as dissemelhanças entre entidades então os gráficos visualizam correctamente as dissemelhanças.

A figura 4.13 mostra a MST para a matriz de dissemelhanças calculada com a medida de distâncias euclideanas ponderadas sobreposta à respectiva análise de MDSC. Verifica-se que a situação da Nigéria, Egipto, Áustria e Finlândia constituem distorções da análise, ou seja, distorções na redução do número de dimensões.

4.2. "*MULTIDIMENSIONAL SCALING*" ORDINAL

O MDSO propõe uma configuração de pontos na qual a ordenação das distâncias entre os pontos se ajusta à ordenação das dissemelhanças entre objectos, mesmo que existam grandes

FIGURA 4.7 – MDSC para 30 paises de Africa e Europa
(δ_{ij}=distancia euclideana ponderada)

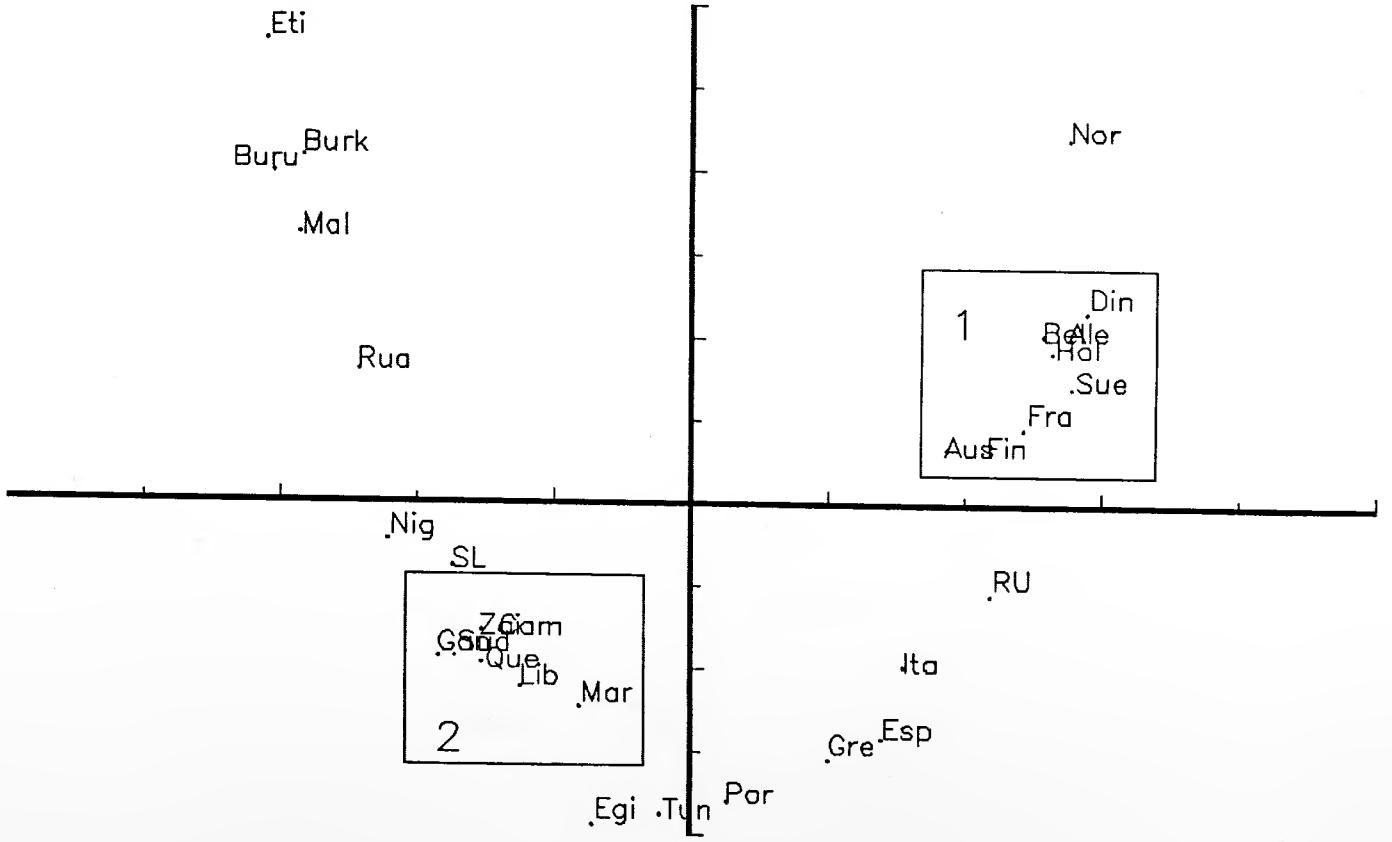


FIGURA 4.8 – MDSC para 30 paises de Africa e Europa
(δ=distancia euclideana ponderada) – Sector 1

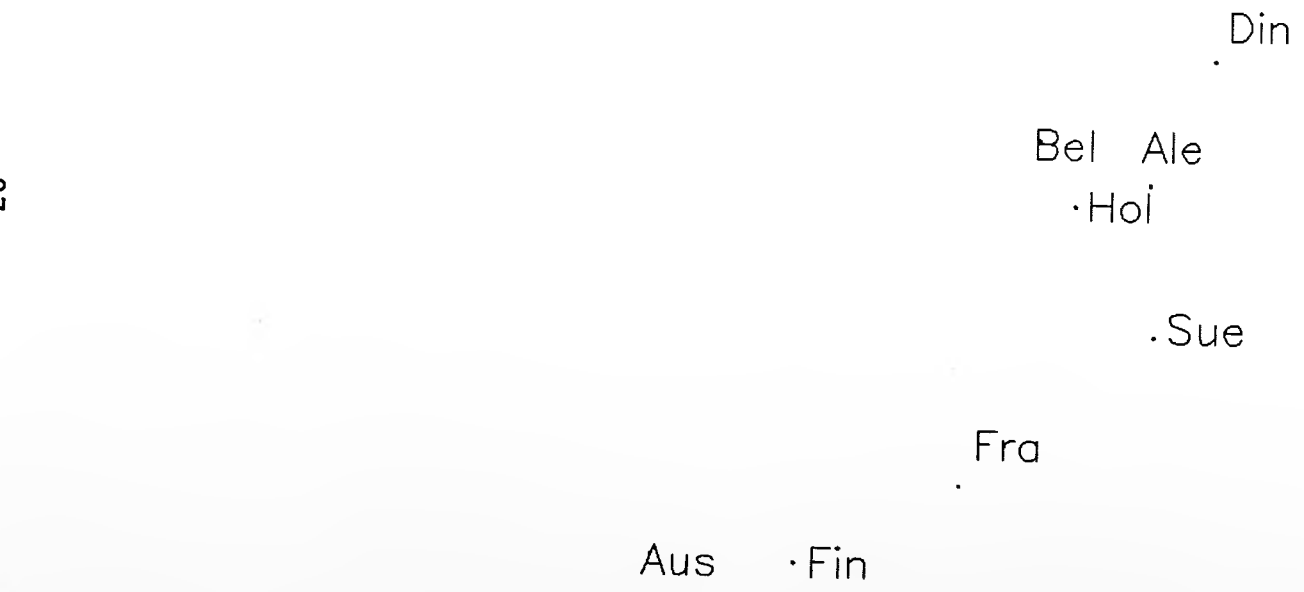


FIGURA 4.9 – MDSC para 30 paises de Africa e Europa (δ_{ij}=distancia euclideana ponderada) – Sector 2

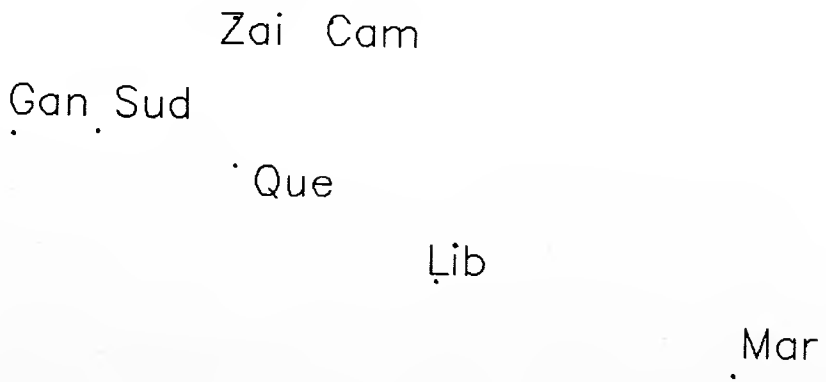


FIGURA 4.10 – MDSC para 30 paises de Africa e Europa
(δ_{ij}=distancia "*city block*" ponderada)

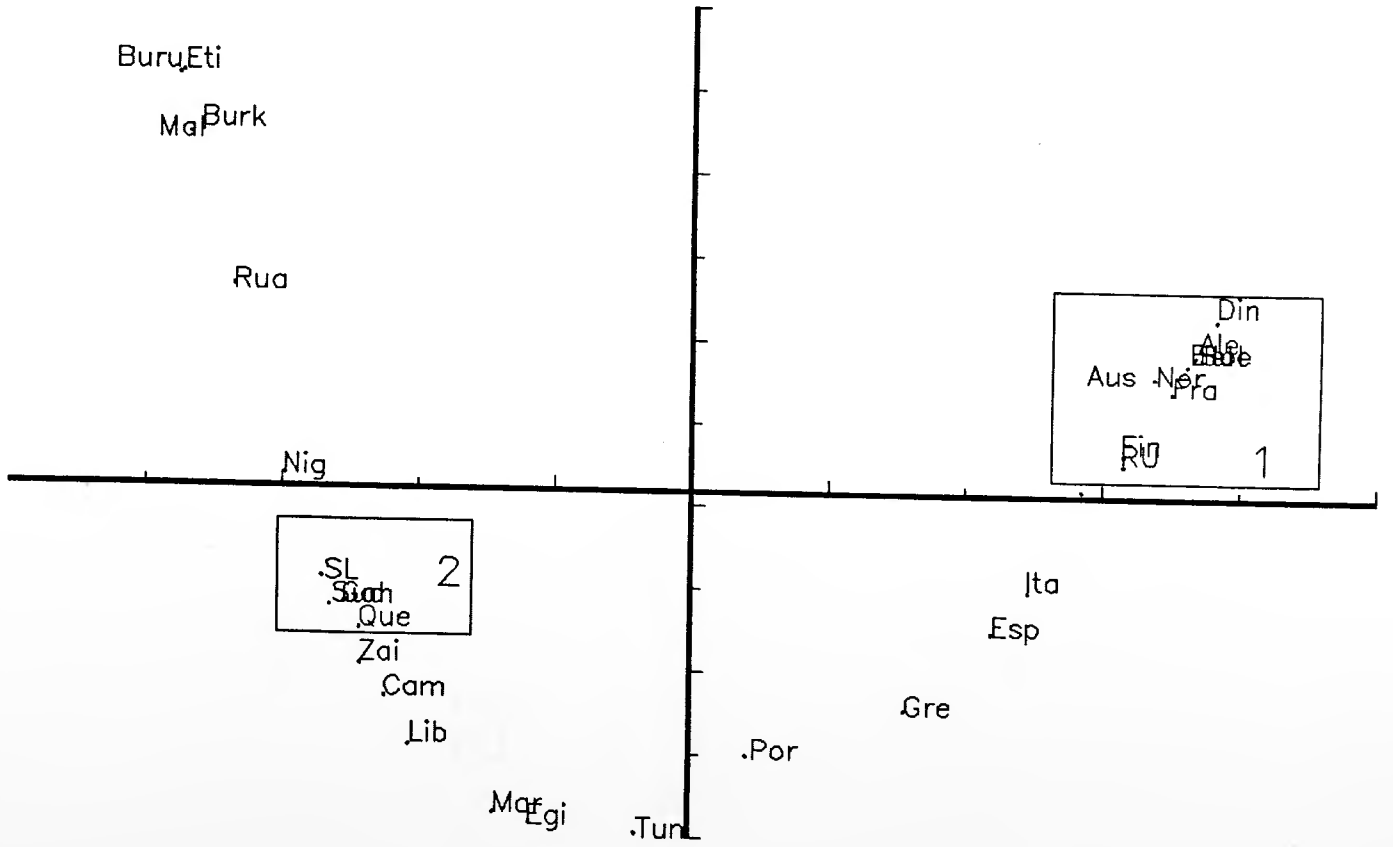


FIGURA 4.11 – MDSC para 30 paises de Africa e Europa (δ_{ij}=distancia "*city block*" ponderada) – Sector 1

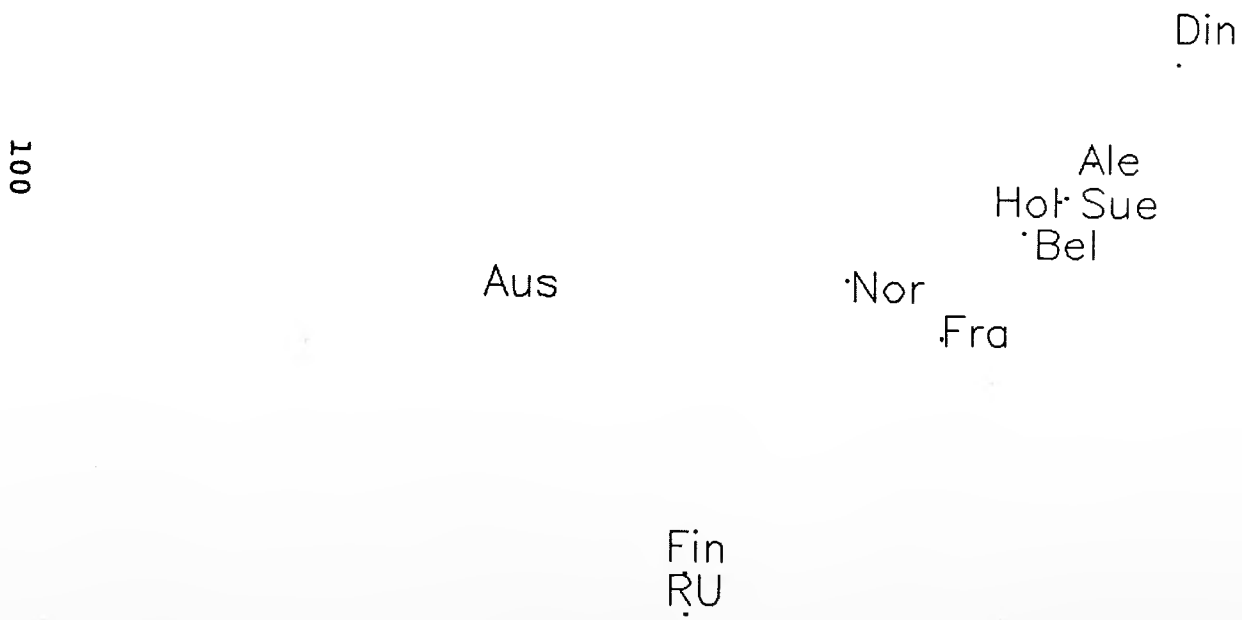


FIGURA 4.12 – MDSC para 30 paises de Africa e Europa (δ_{ij}=distancia "*city block*" ponderada) – Sector 2

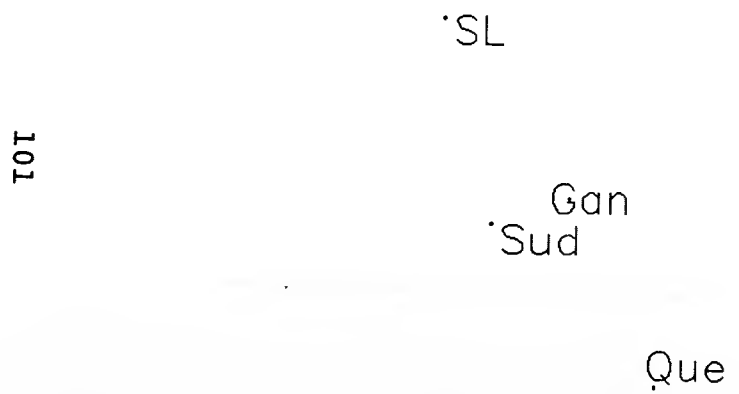
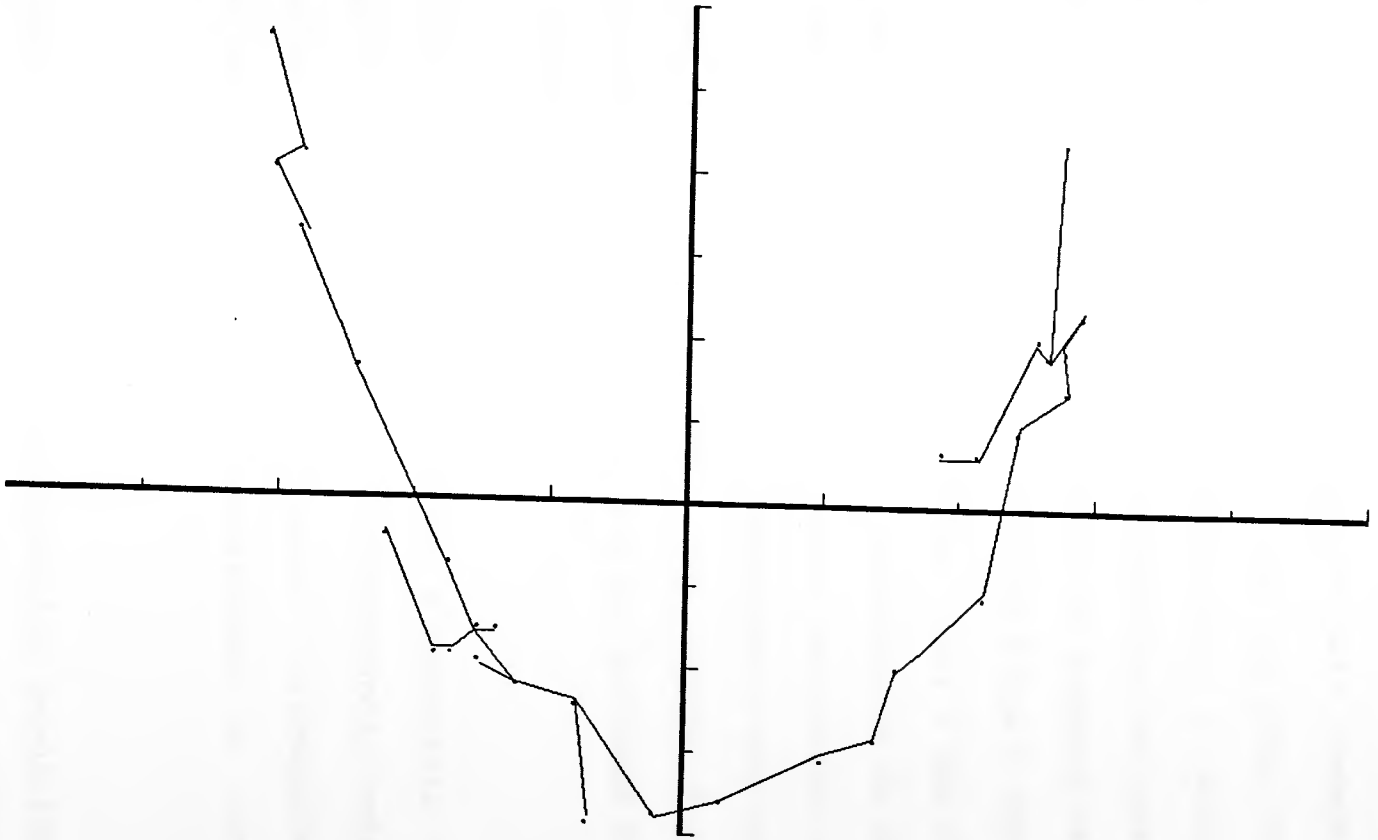


FIGURA 4.13 – MST da matriz de distancias euclideanas ponderadas



discrepâncias entre alguns dos verdadeiros valores. Tal como no MDSC é desejável representar os dados sem grande distorção em duas dimensões. Em muitas ocasiões os dados são obtidos numa forma ordinal, de modo que tudo o que existe para começar são ordenações na matriz de dissemelhanças em vez dos seus valores numéricos. Por exemplo numa investigação sobre o sabor de determinadas bebidas, uma indivíduo pode deparar-se com pares de bebidas não identificadas e ser-lhe pedido que as compare segundo o gosto. Tudo o que o sujeito consegue expressar é que a bebida A é mais parecida com a bebida **B** do que a bebida C ou, o que é mais importante, que as bebidas A e B são menos semelhantes do que as bebidas C e D e assim por diante. Juntando esta informação obtém-se uma matriz de dissemelhanças contendo ordenações e não valores numéricos. Uma outra situação de uso duma escala ordinal é aquela em que os valores numéricos foram obtidos mas por qualquer motivo a sua precisão é suspeita.

As distâncias entre pontos da configuração a construir devem estar monotonamente relacionadas com as dissemelhanças entre os objectos, ou seja, pequenas dissemelhanças correspondem a pequenas distâncias e grandes dissemelhanças a grandes distâncias.

Como motivação para os conceitos e procedimentos envolvidos na abordagem de Shepard-Kruskal, considere-se o exemplo apresentado por Gnanadesikan(1977). Existem 4 objectos e seis valores de

dissemelhança entre eles. Sendo δ_{ij} o valor da dissemelhança entre os objectos i e j, para i,j=1,2,3,4; eles encontram-se ordenados da seguinte maneira:

$$\delta_{23} < \delta_{12} < \delta_{34} < \delta_{13} < \delta_{24} < \delta_{14}.$$

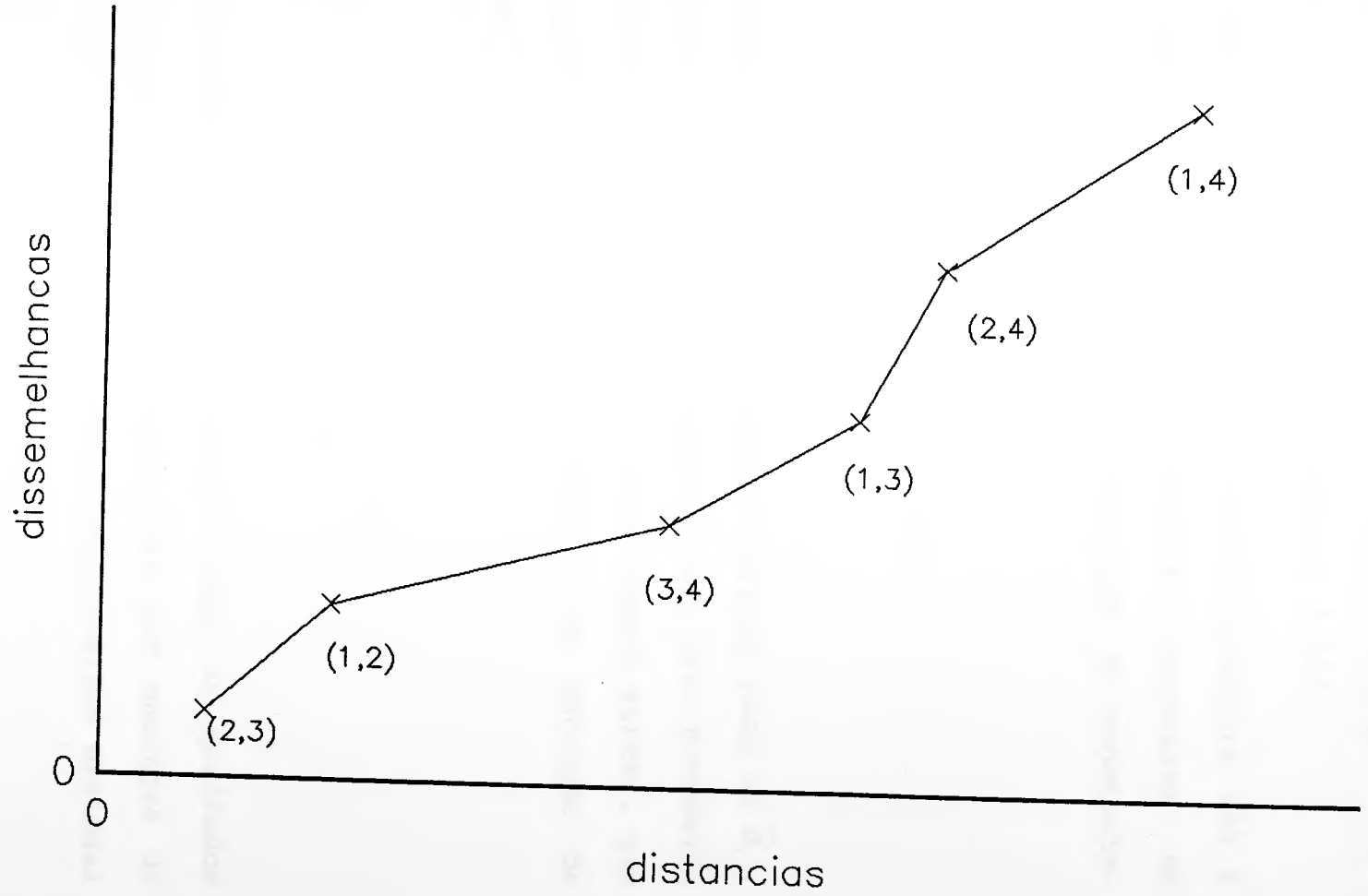
Isto significa que os segundo e terceiro objectos são considerados os menos dissemelhantes, os primeiro e segundo menos dissemelhantes a seguir aos segundo e terceiro, e, por aí adiante, sendo os primeiro e quarto objectos os mais dissemelhantes. Representem-se os objectos num espaço euclideano e calculem-se as distâncias euclideanas d_{ij} entre cada par de pontos: no caso ideal em que a monotonia é perfeita, as distâncias satisfazem a seguinte relação,

$$d_{23} \leq d_{12} \leq d_{34} \leq d_{13} \leq d_{24} \leq d_{14},$$

pois a ordem das distâncias entre os pontos está de acordo com a ordem das dissemelhanças. Assim, se a monotonia é perfeita, a representação gráfica de d_{ij} contra δ_{ij} é tal que os segmentos de recta que unem os pontos progridem simultaneamente para cima e da esquerda para a direita (figura 4.14). Porém se as distâncias não se comportam desta maneira, mas se tem, por exemplo,

$$d_{23} \leq d_{34} \leq d_{12} \leq d_{13} \leq d_{14} \leq d_{24},$$

FIGURA 4.14 – Relacao entre dissemelhancas e distancias satisfazendo a condicao de monotonia



a monotonia é violada. A linha que une os pontos desloca-se para cima e da esquerda para a direita nalguns locais mas, da direita para a esquerda noutros. A cadeia dos segmentos de recta apresenta, agora, um movimento em *zigzag* (figura 4.15).

Para os frequentes casos em a propriedade de monotonia não é completamente satisfeita torna-se necessário encontrar um conjunto de valores $\hat{d}_{ij}$ que satisfaçam a restrição de monotonia. No exemplo,

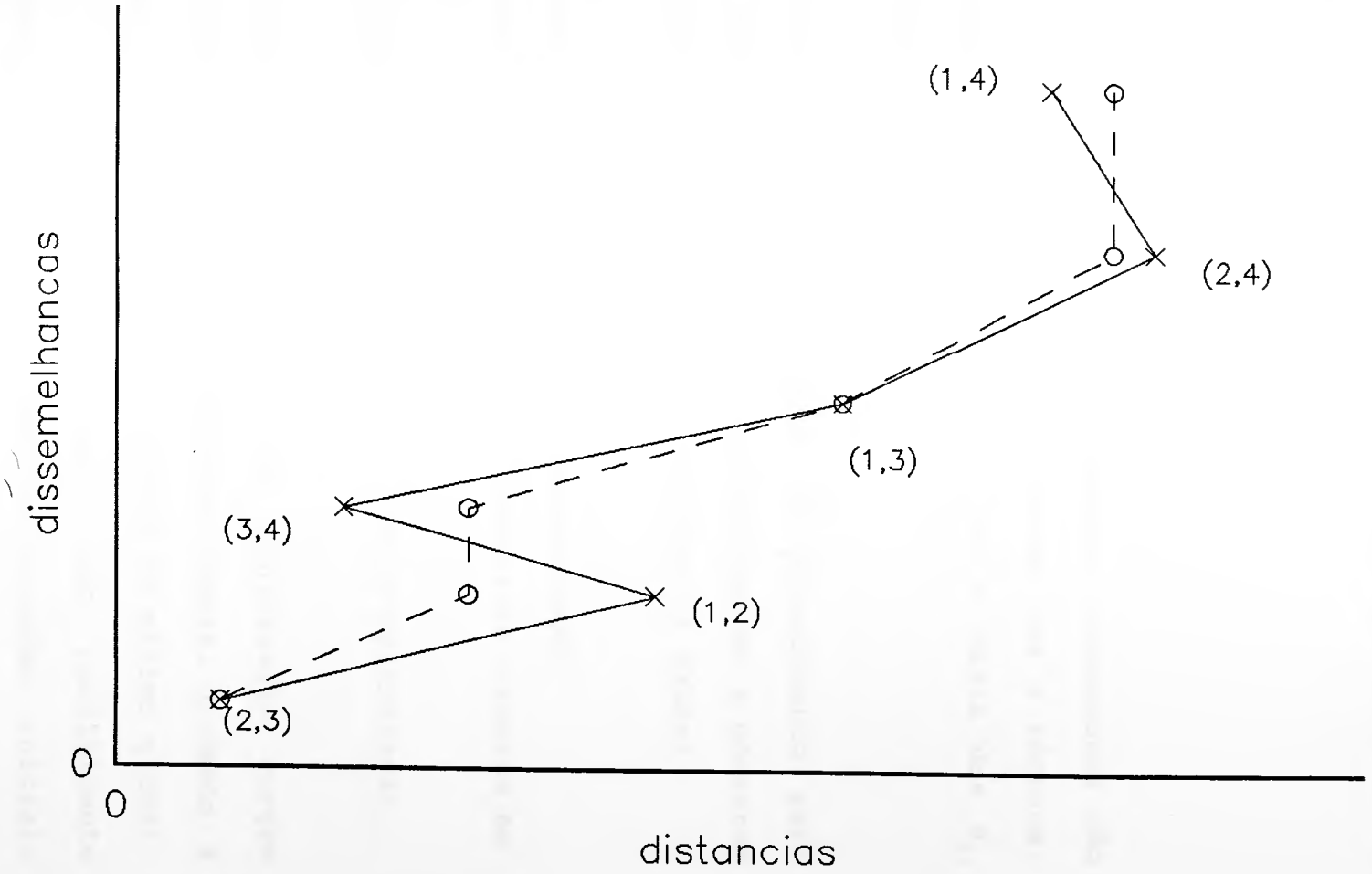
$$\hat{d}_{23} \leq \hat{d}_{12} \leq \hat{d}_{34} \leq \hat{d}_{13} \leq \hat{d}_{24} \leq \hat{d}_{14}.$$

Não é possível determinar nenhuma solução analítica para os $\hat{d}_{ij}$, pelo que se adopta um procedimento iterativo. Com este propósito Kruskal definiu uma função objectivo S, a que chamou ***Stress***, que é usada como base para o método sistemático de obtenção de valores ajustados $\hat{d}_{ij}$,

(4.13) $$S = \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2 \; / \; \sum_{i<j} d_{ij}^2$$

Repare-se que o ***Stress*** não é mais que uma soma de quadrados residual normalizada de modo a ser invariante sob mudanças de escala. Aliás o nome é uma abreviatura de "***STandardized REsidual Sum of Squares***".

FIGURA 4.15 – Relacao entre dissemelhancas e distancias violando a condicao de monotonia



Dado um conjunto de d_{ij}, podem escolher-se os $\hat{d}_{ij}$ de modo a minimizar S, sujeito à restrição de monotonia. Este problema de minimização é equivalente a um problema de regressão de mínimos quadrados monótona entre d_{ij} e δ_{ij},

(4.14) $$d_{ij} = f(\delta_{ij}) + \varepsilon_{ij}$$

em que $\hat{d}_{ij} = f(\delta_{ij})$, sendo f uma função monótona. Pormenores não são importantes aqui; no entanto, pode dizer-se que a técnica, *grosso modo*, envolve a escolha de $\hat{d}_{ij}$ que são a média dos d_{ij} para cada bloco de não monotonia.

O mesmo autor (Kruskal,1964) descreveu um procedimento para optimizar esta função que, ainda hoje, é de grande uso e conserva no essencial a forma como foi definido há 30 anos. A saber:

1. Escolher a dimensão t desejada para a representação.
2. Seleccionar uma configuração inicial arbitrária de n pontos em t dimensões.
3. Minimizar a função S utilizando um algoritmo de optimização.

Tal como com todos os procedimentos de minimização surgem problemas por causa da presença de mínimos locais, podendo a solução convergir para um mínimo local em vez de mínimo global. Este problema é irresolúvel mas pode ser parcialmente ultrapassado repetindo o passo 3. com configurações iniciais

diferentes.

Para determinar a melhor dimensão t na qual a amostra é representada há que acrescentar dois passos:

4. Repetir todo o procedimento para um conjunto m de valores de t, $t_1 \leq t \leq t_m$ (na prática m=5). A melhor estratégia para a sua realização é :

   i) Tomar o maior t ($t=t_m$) e efectuar os passos 1. a 4..

   ii) Considerar as coordenadas dos pontos nas primeiras t-1 dimensões que funciona como a configuração inicial para o valor de t seguinte ($t_{m-1}=t_m-1$).

   iii) Repetir ii) para todos os valores de t.

5. Fazer um gráfico ('*scree graph*') com os valores de S contra os valores de t e procurar no gráfico um ângulo onde S diminui relativamente devagar. O valor da abcissa é um bom valor para t.

Decidir o número de coordenadas da configuração é, como afirma, Everitt(1983),

> *"muito mais uma questão de substância que uma questão estatística".*

Mesmo que existisse um método estatístico razoável para

determinar o número correcto de dimensões, tal não seria suficiente visto que a solução a duas dimensões é, na prática, a de maior importância.

A função *Stress* serve, ainda, como medida da bondade do ajustamento da configuração de pontos que resulta como solução para um dado t estando compreendida no intervalo [0,1]:

Stress	≥ 0.2	≈ 0.1	≈ 0.05	≈ 0.25	0
Ajustamento	pobre	suficiente	bom	excelente	perfeito

Porém estas indicações são demasiado simplistas porque S depende do número de objectos n e do número de dimensões t. O Stress diminui quando n diminui e/ou t aumenta. Como seria de esperar ao aumentar t obtêm-se melhores representações pelo que S diminui. O investigador deve encontrar o menor t para o qual é possível obter um bom valor de S. Assim a quantidade de *Stress* que se está preparado para tolerar é uma decisão arbitrária.

Outra função S muito usada e designada por *Sstress* substitui d_{ij} e $\hat{d}_{ij}$ por d^2_{ij} e $\hat{d}^2_{ij}$,

$$Ss = \sum_{i<j} (d^2_{ij} - \hat{d}^2_{ij})^2 \; / \; \sum_{i<j} d^4_{ij} \; . \tag{4.15}$$

Uma medida que também pode ser usada para avaliar o ajustamento

entre d_{ij} e $\hat{d}_{ij}$ é o coeficiente G_1 definido pela expressão (4.10). Valores de G_1 próximos de um correspondem a valores de *S* ou *Ss* próximos de zero.

É possível, à semelhança do que acontece com o MDSC, efectuar uma rotação à configuração mudando drasticamente as projecções sem que as distâncias entre os pontos se alterem. O problema da orientação não é, mais uma vez, unicamente determinado.

Chatfield(1980) relata que várias experiências para comparar MDSC e MDSO foram levadas a cabo na Universidade de Bath. As experiências começavam com uma configuração de pontos conhecida e, de seguida, as distâncias euclideanas entre os diversos pares de pontos eram sujeitas a uma variedade de transformações e erros aleatórios, reconstruindo-se, por fim, a configuração inicial usando as duas técnicas. Concluíram os investigadores que se obtinham mais ou menos os mesmos resultados quando as distâncias eram aproximadamente euclideanas mas quando não eram o "*scaling*" ordinal gerava melhores resultados.

Para ilustrar o método descrito pediu-se a 20 estudantes de Economia, com idades entre os 18 e 22 anos, que manifestassem a sua opinião àcerca do nível socio-económico dos 30 países do Exemplo. Os alunos compararam pares de países usando uma escala de pontuações de 1 a 9 - 1 significa perfeita semelhança e 9 máxima diferença -. Foi calculada a matriz da média das opiniões

à qual se aplicou MDSO e o resultado consta da figura 4.16. O sector assinalado está ampliado na figura 4.17.

O ajustamento da configuração à matriz de dissemelhanças é de boa qualidade traduzido pelos valores das medidas G_1=0,953, *Stress*=0,119 e *Sstress*=0,120. Da observação da figura 4.16 é possível aperceber a tendência dos estudantes para extremar posições Europa *versus* África. Os estudantes distinguem entre os países europeus o núcleo de países do Norte a que juntam a Itália e afastam deste grupo, por um lado, a Espanha e Grécia e, por outro, a Finlândia e Portugal. No seio dos países africanos é menor a sensibilidade para a constituição de grupos pelo que os países surgem todos mais ou menos próximos dos outros, salvo os extremos Etiópia e Nigéria.

FIGURA 4.16 – MDSO para 30 paises de Africa e Europa

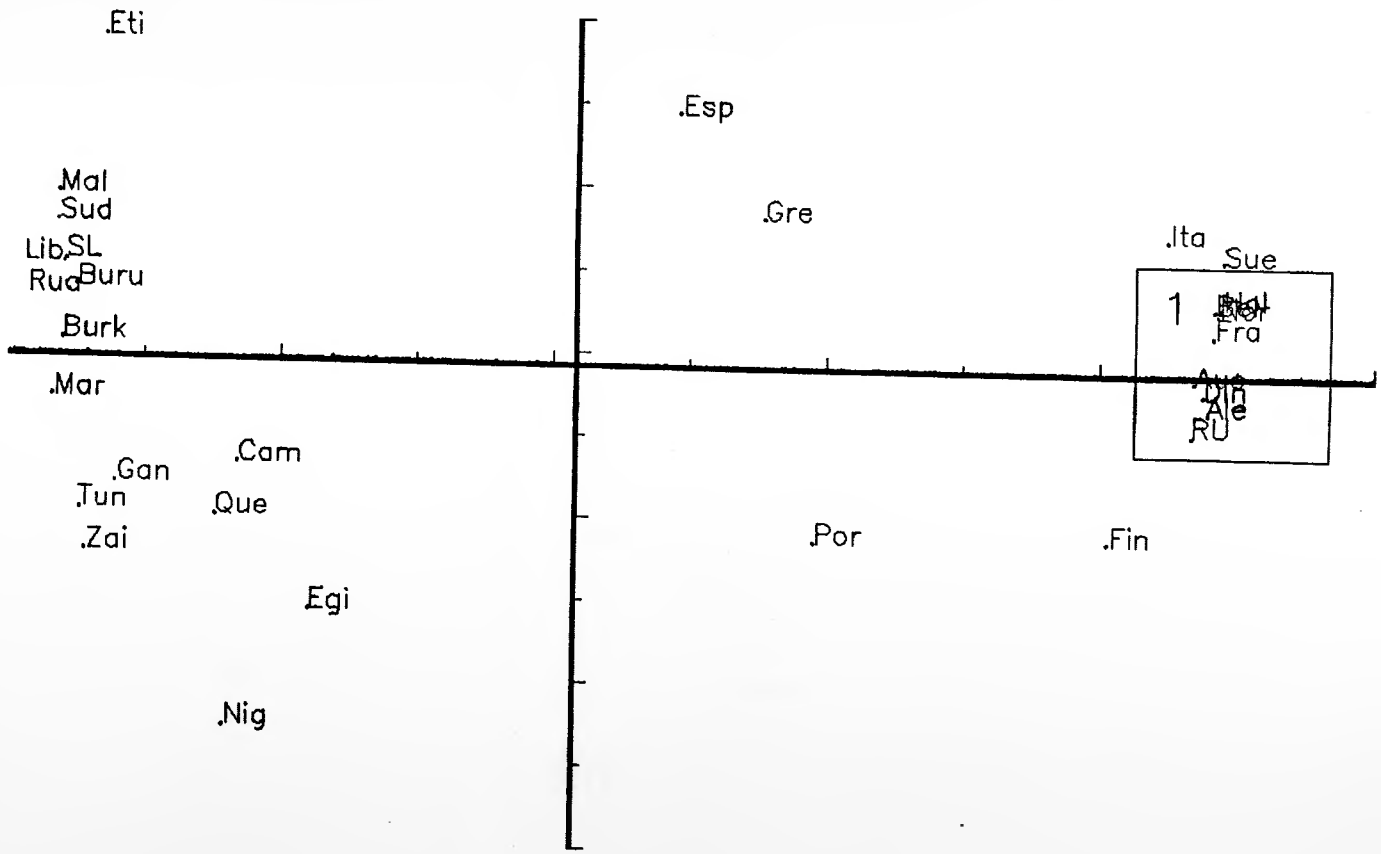
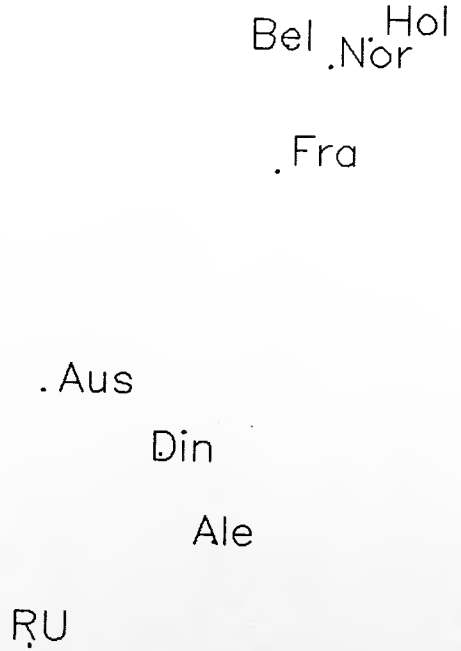


FIGURA 4.17 – MDSO para 30 paises de Africa e Europa
Sector 1



CAPÍTULO 5

ESQUEMA DE REPRESENTAÇÃO BILATERAL: *"BIPLOT"*

O capítulo 3 dá conta de métodos apropriados para construir representações das linhas ou das colunas de uma matriz de observações **X**. A configuração resultante situa-se, quer no espaço das variáveis, quer no espaço dos sujeitos, conforme as entidades que se desejam estudar. Por vezes revela-se interessante extrair informação da representação conjunta de ambos os espaços que, por esse motivo, se denomina representação bilateral.

Gabriel(1971) analisou um conjunto de dados relativos à percentagem de habitações possuindo diversas características tais como casa de banho, água canalizada, electricidade, etc., em diferentes bairros de certas cidades de Israel. Da leitura de uma configuração conjunta das localidades e das características ressalta não só a semelhança entre localidades (traduzida pela proximidade dos pontos representativos das localidades) mas também as características mais importantes partilhadas pelos grupos de cidades semelhantes.

Krzanoswki(1990) relata mais exemplos da utilidade da representação conjunta.

> *"Num estudo de mercado sobre o padrão de compra de produtos de consumo entre vários estratos da população seria útil uma representação que mostrasse as relações entre os estratos e caracterizasse, ainda, cada agrupamento por tipo de produto. Num estudo sociológico àcerca das qualidades pessoais atribuídas a várias profissões seria importante salientar as semelhanças entre profissões e também quais as qualidades pessoais características de cada grupo de profissões."*

"*Biplot*" é um método que se insere na abordagem de projeçcão e optimização e que permite a representação simultânea dos indivíduos e das variáveis, "lendo", ao mesmo tempo, as duas entidades da matriz **X**. Foi Gabriel(1971) quem descreveu o procedimento para construir o "*biplot*", fazendo diversas aplicações - Gabriel(1972). O mesmo autor utilizou o "*biplot*" como instrumento de diagnóstico da modelização dos dados multivariados - Bradu e Gabriel(1978).

Nas Ciências Sociais, sobretudo nas áreas de Psicologia e Sociologia, tornou-se muito popular a análise de matrizes que consistem em classificações ou ordenações de um conjunto de *p*

estímulos feito por cada um dos n sujeitos arranjados de tal forma que as linhas da matriz correspondem aos sujeitos e as colunas aos estímulos. O procedimento de representação bilateral deste tipo de matrizes é, em tudo, equivalente ao "*biplot*" apesar da técnica ser conhecida pela nome de "*preference scaling*" ou "*repertory grid analysis*".

O "*biplot*" baseia-se no resultado que factoriza qualquer matriz $\mathbf{X}$ $(n\times p)$ de característica k $(k\leq\min(n,p))$ na forma,

$$\mathbf{X} = \mathbf{G}\mathbf{H}^T, \tag{5.1}^*$$

em que $\mathbf{G}$ é matriz $(n\times k)$ e $\mathbf{H}$ é matriz $(p\times k)$, ambas necessariamente de característica k.

A factorização (5.1) reescreve-se do seguinte modo,

$$x_{ij} = \mathbf{g}_i^T\mathbf{h}_j \ , \qquad (i=1,\ldots,n;j=1,\ldots,p), \tag{5.2}$$

onde x_{ij} é o elemento genérico de $\mathbf{X}$, $\mathbf{g}_i^T$ é a i-ésima linha de $\mathbf{G}$ e $\mathbf{h}_j$ é a j-ésima coluna de $\mathbf{H}$. Os vectores $\mathbf{g}_1,\ldots,\mathbf{g}_n$ são designados marcadores de linha de $\mathbf{X}$ e os vectores $\mathbf{h}_1,\ldots,\mathbf{h}_p$ marcadores de coluna. Cada elemento de $\mathbf{X}$ é, assim, expresso como o produto interno entre o marcador da linha e o marcador da coluna correspondentes. Visto que cada um destes vectores tem dimensão k, (5.2) permite a representação da matriz $\mathbf{X}$ através de $n+p$

vectores num espaço k-dimensional.

Quando a característica de $\mathbf{X}$ é dois, os vectores $\mathbf{g}_i$ e $\mathbf{h}_j$ têm ambos dimensão dois e podem ser configurados no plano. É a este gráfico que se dá o nome de "*biplot*". Gabriel(1981) esclarece que o prefixo "bi" não significa tratar-se de uma representação bidimensional mas da representação conjunta das linhas e colunas de uma matriz.

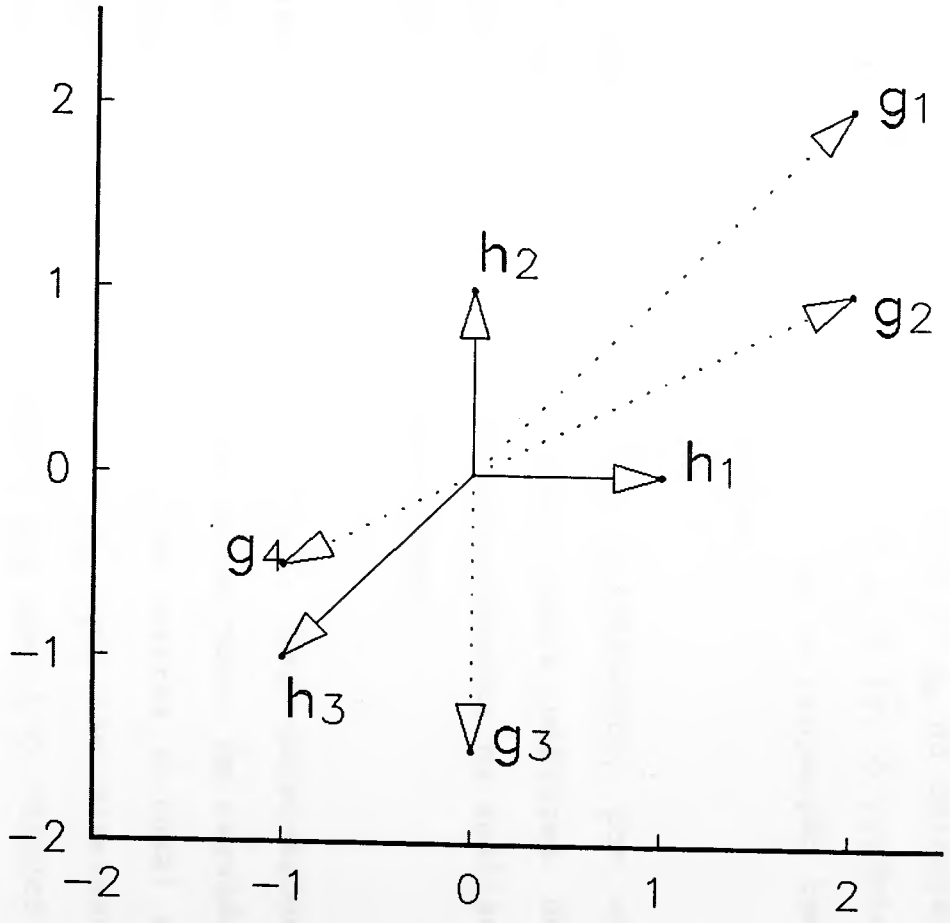
O produto interno de dois vectores é visualizado como o produto do comprimento de um dos vectores pelo comprimento da projecção do outro vector no primeiro. O "*biplot*" transmite, assim, rapidamente a estrutura da matriz $\mathbf{X}$, podendo, por exemplo, detectar quais as linhas ou colunas que são proporcionais (marcadores com a mesma direcção) ou quais os elementos nulos da matriz (ortogonalidade entre marcadores).

Considere-se um exemplo simples. Seja a matriz $\mathbf{X}$ (4x3) factorizada da forma (5.1) como segue:

$$\begin{bmatrix} 2 & 2 & -4 \\ 2 & 1 & -3 \\ 0 & -3/2 & 3/2 \\ -1 & -1/2 & 3/2 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 1 \\ 0 & -3/2 \\ -1 & -1/2 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

e o respectivo "*biplot*" desenhado na figura 5.1. Os marcadores das linhas 2 e 4 têm a mesma direcção embora sentidos contrários

FIGURA 5.1 – BIPLOT da matriz X



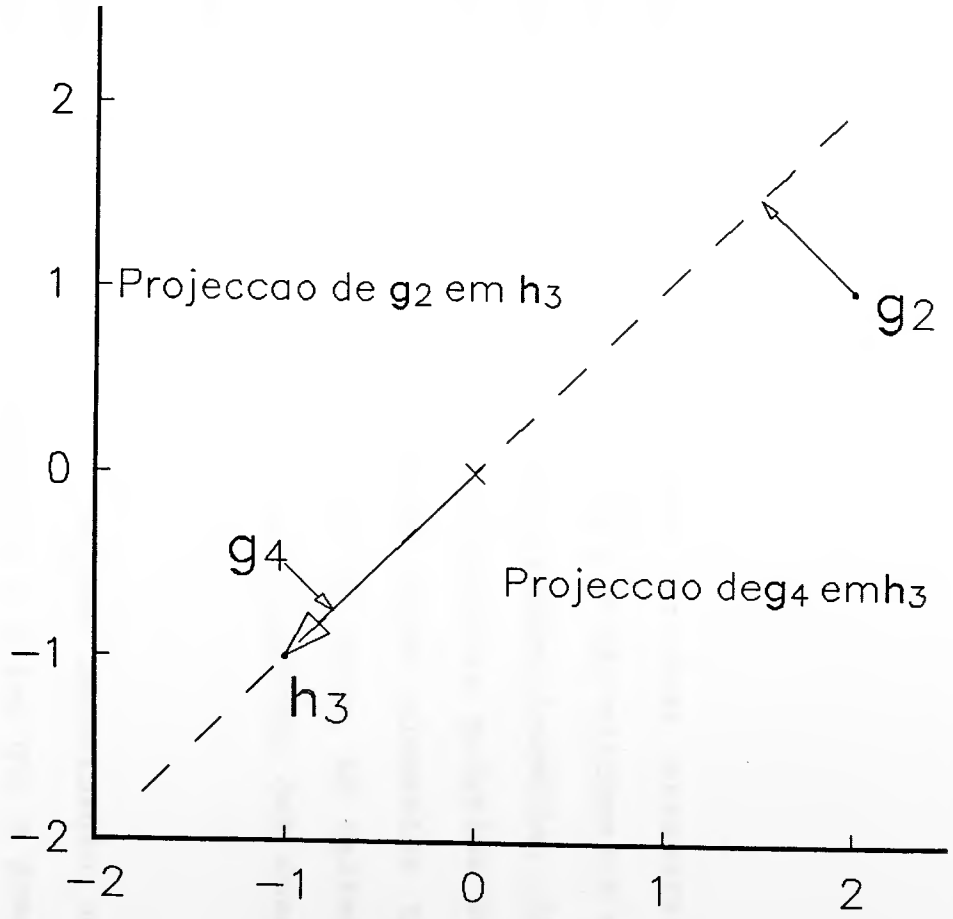
significando que existe uma proporcionalidade inversa entre as duas linhas. O marcador da coluna 1 e o marcador da linha 3 formam um ângulo recto entre si pelo que o elemento x_{31} é nulo.

O interpretação do "*biplot*" através do conceito de produto interno é feito a partir da figura 5.2 onde estão representados os elementos $x_{2,3}$ e $x_{4,3}$ como o produto interno de, respectivamente, $\mathbf{g}_2$ e $\mathbf{h}_3$ e $\mathbf{g}_4$ e $\mathbf{h}_3$. A projecção de $\mathbf{g}_2$ na direcção de $\mathbf{h}_3$ tem comprimento $3/\sqrt{2}$; o comprimento de $\mathbf{h}_3$ é $\sqrt{2}$; o produto interno é -3 significando o sinal negativo que a projecção tem sentido oposto ao do vector onde é projectado.

A matriz **X** do exemplo foi representada, exactamente, por um "*biplot*" porque tinha característica dois. Para matrizes de característica superior, o que acontece frequentemente na análise multivariada, constrói-se um "*biplot*" aproximado.

Num primeiro passo aproxima-se a matriz **X**, das observações originais, por uma matriz $\mathbf{X}_{[2]}$ de característica dois. De seguida factoriza-se $\mathbf{X}_{[2]}$ como o produto (5.1) de uma matriz **G** (nx2) e uma matriz $\mathbf{H}^T(2xp)$. Com os vectores $\mathbf{g}_i$ e $\mathbf{h}_j$, que têm dimensão dois, representa-se o "*biplot*" da matriz $\mathbf{X}_{[2]}$ que é o "*biplot*" aproximado da matriz original **X**. Se **X** fôr satisfatoriamente aproximada por $\mathbf{X}_{[2]}$, o "*biplot*" de $\mathbf{X}_{[2]}$ permitirá uma útil inspecção visual aproximada da própria matriz **X**. Neste caso, os produtos internos dos marcadores de linhas e de colunas serão

Figura 5.2 — Representacao, em termos do produto interno, de dois elementos da matriz X



aproximações dos elementos de **X**.

Para aproximar a matriz rectangular **X** (nxp) de característica k ($k>2$) por uma matriz **X** (nxp) mas de característica dois, Gabriel usa o resultado de Eckart e Young (vidé Anexo-(4.7)). A aproximação é calculada a partir da decomposição da matriz **X** em valores singulares, considerando apenas as duas primeiras colunas das matrizes singulares,

$$\mathbf{X}_{[2]} = \mathbf{U}_2 \, \mathbf{T}_2 \, \mathbf{V}_2^{\mathrm{T}} \, . \tag{5.3}$$

O índice 2 significa que se tomam apenas as duas primeiras colunas das matrizes **U**, **T** e **V**. As matrizes **U** e **V** são ortogonais e os vectores-coluna que as constituem são, respectivamente, os vectores próprios de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$, $\mathbf{u}_i$ (i=1,...,k) e vectores próprios de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\mathbf{v}_j$ (j=1,...,k). A matriz **T** é diagonal cujos elementos τ_s (s=1,...,k) são os valores singulares de **X**, ou seja, as raízes quadradas positivas dos valores próprios não negativos das duas matrizes simétricas $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ e $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

Deste modo garante-se ter sido obtida a melhor aproximação no sentido dos mínimos quadrados, o que equivale a dizer que a soma dos quadrados dos desvios entre os elementos de **X** e os elementos homólogos da matriz $\mathbf{X}_{[2]}$ é mínima.

A qualidade da aproximação mede-se, à semelhança de (4.10),

através de,

$$G = \sum_{s=1}^{2} \tau_s^2 \; / \; \sum_{s=1}^{k} \tau_s^2 \,, \tag{5.4}$$

revelando os valores próximos de um que se está em presença de uma boa aproximação.

O resultado de Eckart e Young não é válido quando surge a necessidade de utilizar ponderações como, por exemplo, na presença de matrizes com "*missing values*" aos quais se atribui ponderação zero. Neste caso, multiplica-se o desvio quadrático $(x_{ij}-x_{[2]ij})^2$ por uma ponderação w_{ij} sendo a matriz **X** aproximada através do método de mínimos quadrados ponderados para o qual Gabriel e Zamir(1979) elaboraram um algoritmo.

A factorização (5.1) não é única. Com efeito, se pós-multiplicarmos **G** por qualquer matriz $\mathbf{R}^T$ (2x2) não singular e **H** pela matriz inversa $\mathbf{R}^{-1}$, a factorização,

$$\mathbf{X} = (\mathbf{G}\mathbf{R}^T)(\mathbf{H}\mathbf{R}^{-1})^T, \tag{5.5}$$

é tão válida quanto (5.1). Considere-se a decomposição em valores singulares da matriz **R**, então,

$$\mathbf{R}^T = \mathbf{V}\ \mathbf{T}\ \mathbf{U}^T \quad ,$$

e, $$\mathbf{R}^{-1} = \mathbf{V}\ \mathbf{T}^{-1}\ \mathbf{U}^{T},$$

onde **V**, **U** e T têm o significado habitual. Ora, a transformação $\mathbf{G} \rightarrow \mathbf{GR}^{T}$ consiste numa rotação dos eixos, mudança de escala ao longo dos novos eixos e uma nova rotação enquanto a transformação $\mathbf{H} \rightarrow \mathbf{HR}^{-1}$ consiste nas mesmas rotações e numa mudança de escala recíproca da primeira. Isto dá uma ideia de como as diferentes factorizações e respectivos "*biplots*" estão relacionados. Para mais detalhes relativos a rotações e mudanças de escala deverá consultar-se o capítulo 7.

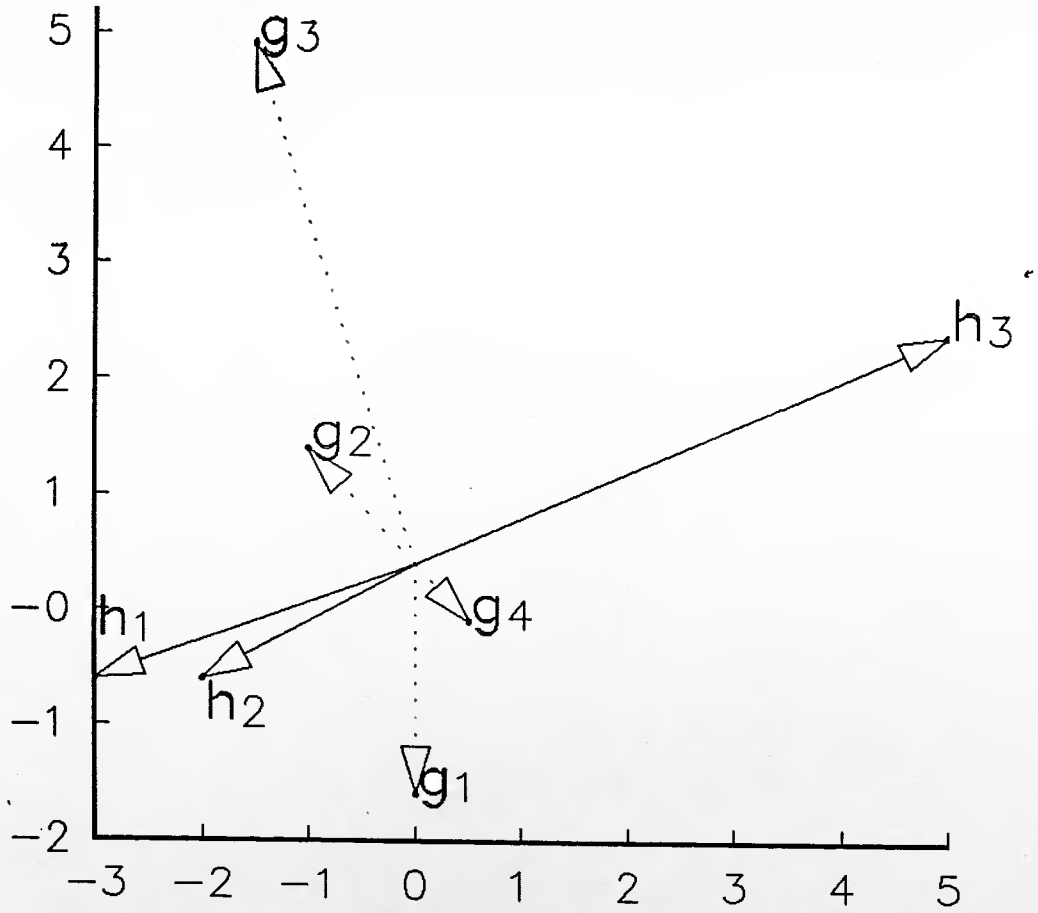
Retome-se o exemplo apresentado algumas linhas atrás e a seguinte matriz $\mathbf{R}^{T}$ expressa pelo produto:

$$\mathbf{R}^{T} = \begin{bmatrix} -1 & 2 \\ 1 & -3 \end{bmatrix}$$

$$= \begin{bmatrix} 0.58 & -0.82 \\ -0.82 & 0.58 \end{bmatrix} \begin{bmatrix} 3.86 & 0 \\ 0 & 0.26 \end{bmatrix} \begin{bmatrix} -0.36 & 0.93 \\ 0.93 & 0.36 \end{bmatrix}.$$

O "*biplot*" que resulta de (5.5) está representado na figura 5.3 e surgiu após efectuar à figura 5.1 uma rotação através de um ângulo de -55.1° (arcsen(-0.82)), reflectir e reescalar por 3.86 e 0.26, respectivamente, as coordenadas dos dois marcadores de linha, reflectir e reescalar por 1/3.86 e 1/0.26, as coordenadas dos dois marcadores de coluna e, finalmente, efectuar uma rotação através de um ângulo de 68.4° (arcsen(0.93)).

FIGURA 5.3 – BIPLOT alternativo da matriz X



Uma vez que existem factorizações alternativas o investigador pode escolher aquela que lhe traga mais vantagens. Se a matriz original não possuir caraterística dois, torna-se necessário encontrar a matriz aproximada $\mathbf{X}_{[2]}$ podendo esta matriz ser factorizada de acordo com a decomposição de $\mathbf{X}$ em valores singulares mas, mesmo assim, existem diversas hipóteses. A partir de (5.3) pode escolher-se,

(5.6)
$$\mathbf{G} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix} \begin{bmatrix} \sqrt{\tau_1} & 0 \\ 0 & \sqrt{\tau_2} \end{bmatrix},$$

$$\mathbf{H}^T = \begin{bmatrix} \sqrt{\tau_1} & 0 \\ 0 & \sqrt{\tau_2} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{bmatrix},$$

ou,

(5.7)
$$\mathbf{G} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix} \begin{bmatrix} \tau_1 & 0 \\ 0 & \tau_2 \end{bmatrix},$$

$$\mathbf{H}^T = \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{bmatrix}.$$

Com a factorização (5.7),

$$\mathbf{H}^T\mathbf{H} = \mathbf{I},$$

o que implica,

(5.8) $$\mathbf{X}\mathbf{X}^T = \mathbf{G}\mathbf{G}^T.$$

A igualdade (5.8) significa que as relações entre as linhas de **X**, ou seja, entre os indivíduos, são representadas pelas correspondentes relações dos marcadores de linha. Para quaisquer duas linhas *i* e *e* de **X** verifica-se que:

a) O produto interno entre as linhas *i* e *e* é igual ao produto interno entre os marcadores das linhas referidas,

$$\mathbf{x}_i^T\mathbf{x}_e = \mathbf{g}_i^T\mathbf{g}_e \ .$$

b) O comprimento da linha *i* é igual ao comprimento do seu marcador,

$$\|\mathbf{x}_i\| = \|\mathbf{g}_i\| \ .$$

c) O coseno do ângulo formado pelas linhas *i* e *e* é igual ao coseno do ângulo formado pelos respectivos marcadores,

$$\cos\ (\mathbf{x}_i\mathbf{x}_e) = \cos\ (\mathbf{g}_i\mathbf{g}_e) \ .$$

d) A distância entre as linhas *i* e *e* é igual à distância entre os seus marcadores,

$$\|\mathbf{x}_i - \mathbf{x}_e\| = \|\mathbf{g}_i - \mathbf{g}_e\| \ .$$

Em alternativa $\mathbf{X}_{[2]}$ pode ser factorizado como,

(5.9)
$$\mathbf{G} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix} ,$$

$$\mathbf{H}^T = \begin{bmatrix} \tau_1 & 0 \\ 0 & \tau_2 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{bmatrix} ,$$

o que significa,

$$\mathbf{G}^T\mathbf{G} = \mathbf{I} \ ,$$

e então,

(5.10)
$$\mathbf{X}^T\mathbf{X} = \mathbf{H}\mathbf{H}^T \ .$$

A expressão (5.10) revela que as relações entre as colunas de $\mathbf{X}$, ou seja, entre as variáveis, são, agora, reproduzidas pelas correspondentes relações dos marcadores de coluna. O que ficou estabelecido nas alíneas a) a d) anteriores verifica-se ao substituir a palavra 'linha' por 'coluna'.

É hábito construir um tipo especial de "*biplot*" utilizando a matriz de observações centrada, $\mathbf{X}^C$, que resulta de retirar a cada elemento da matriz original a média da coluna respectiva,

(5.11) $$\sum_{i=1}^{n} x_{ij}^{c} = 0 \qquad (j=1,\ldots,p) \ .$$

Para a matriz aproximada, de característica dois, $\mathbf{X}_{[2]}^{c}$ escolhe-se a factorização,

(5.12) $$\mathbf{G} = \left[\mathbf{u}_1 \ \mathbf{u}_2\right] \sqrt{n-1} \ ,$$

$$\mathbf{H}^{T} = \frac{1}{\sqrt{n-1}} \begin{bmatrix} \tau_1 & 0 \\ 0 & \tau_2 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^{T} \\ \mathbf{v}_2^{T} \end{bmatrix} .$$

A expressão (5.12) é idêntica, à parte uma constante, a (5.9),

(5.13) $$(\mathbf{X}^{C})^{T}\mathbf{X}^{C} = (n-1)\ \mathbf{H}\mathbf{H}^{T} \ ,$$

porque,

$$\mathbf{G}^{T}\mathbf{G} = (n-1)\ \mathbf{I} \ ,$$

o que equivale a evidenciar as relações entre colunas através das relações entre os marcadores.

Uma vez que a matriz estimada de variâncias-covariâncias das p variáveis se define,

(5.14) $$\mathbf{S} = \frac{1}{n-1} (\mathbf{X}^C)^T \mathbf{X}^C ,$$

a factorização (5.12) permite escrever,

(5.15) $$\mathbf{H}\mathbf{H}^T = \mathbf{S} ,$$

e $$\mathbf{G}^T\mathbf{G} = \mathbf{X}^C \mathbf{S}^{-1} (\mathbf{X}^C)^T .$$

Neste caso o "*biplot*" é apropriado para inspectionar as variâncias, covariâncias e correlações entre variáveis ao examinar o comprimento dos marcadores de coluna, o ângulo e o produto interno entre eles e, ainda, analisar as diferenças entre os sujeitos. Assim,

a) A covariância entre as variáveis j e l é aproximada pelo produto interno,

$$\mathbf{s}_j^T\mathbf{s}_l \approx \mathbf{h}_j^T\mathbf{h}_l .$$

b) O desvio-padrão da j-ésima variável pelo comprimento,

$$\|\mathbf{s}_j\| \approx \|\mathbf{h}_j\| .$$

c) A correlação entre as j e l-ésimas variáveis por,

$$\cos(\mathbf{s}_j\mathbf{s}_l) \approx \cos(\mathbf{h}_j\mathbf{h}_l) .$$

d) A distância de Mahalanobis entre os sujeitos *i* e *e* é aproximada por,

$$(\mathbf{x}_i - \mathbf{x}_e)\, \mathbf{S}^{-1}\, (\mathbf{x}_i - \mathbf{x}_e)^T \approx \|\mathbf{g}_i - \mathbf{g}_e\|^2 .$$

A qualidade da aproximação mede-se, em substituição de (5.4), através de,

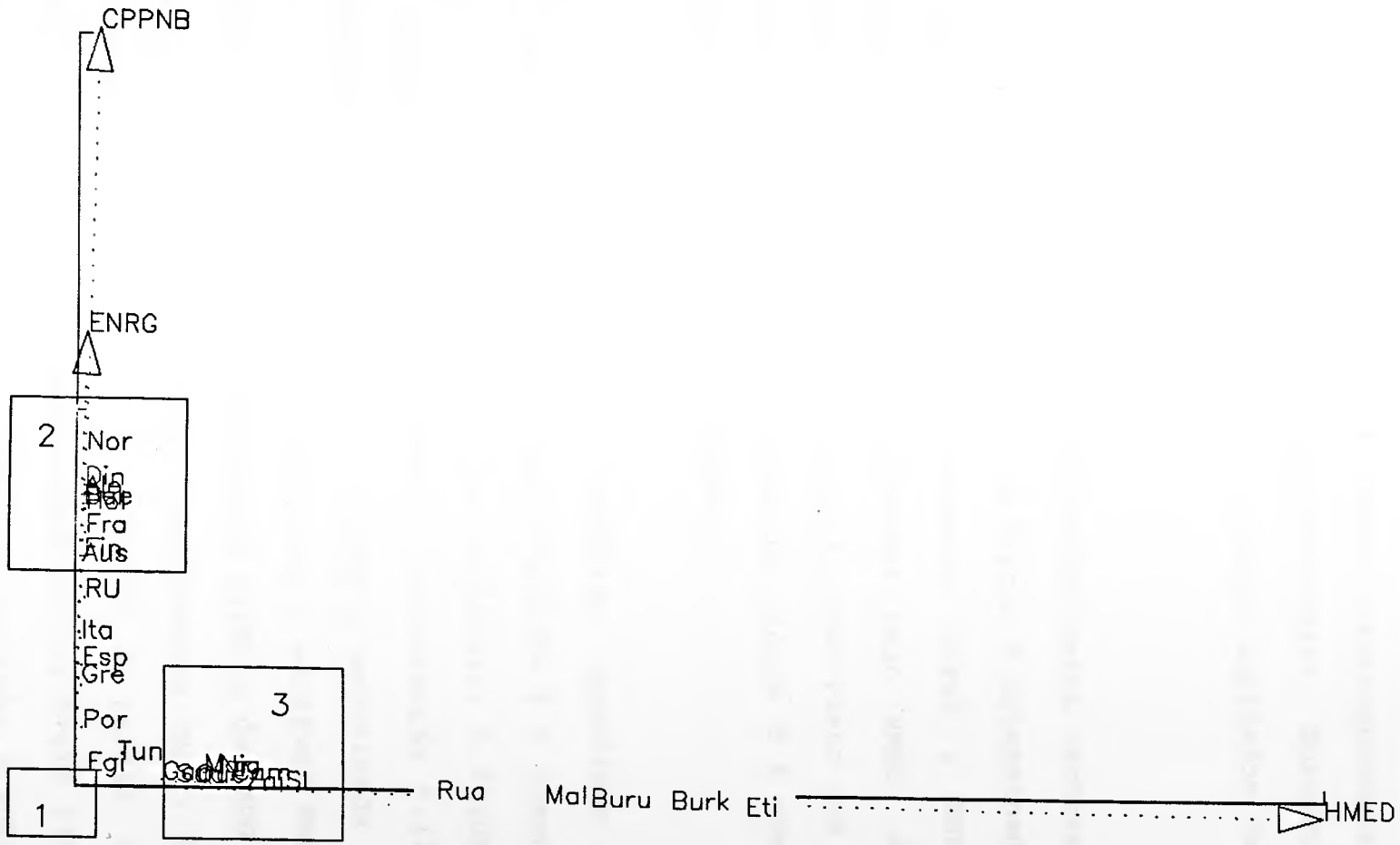
$$(5.16)^* \qquad G_S = \sum_{j=1}^{2} \tau_j^4 \Big/ \sum_{j=1}^{k} \tau_j^4 .$$

Por vezes é útil considerar só a configuração relativa às variáveis desenhando-se, no gráfico, apenas os marcadores de colunas. Corsten e Gabriel(1976) chamaram a esta representação "*h-plot*".

Uma última recomendação para tornar os "*biplots*" menos confusos aconselha representar as colunas por setas e as linhas por pontos.

A figura 5.4 retrata o "*biplot*" do Exemplo. O ajustamento da matriz aproximada de característica dois à matriz original foi obtido com uma qualidade de G = 0,79. Da leitura da figura 5.4

FIGURA 5.4 – "*BIPLOT*" para 30 paises de Africa e Europa



constata-se a existência dos agrupamentos habituais: países do Norte da Europa por um lado, países do Sul da Europa aos quais se juntam a Tunísia e o Egipto e países africanos, distinguindo-se no seio destes o grupo dos infelizes BBERM(Burkina, Burundi, Etiópia, Ruanda e Malawi). Os sectores 2 e 3 estão ampliados nas figuras 5.6 e 5.7.

A localização dos países no gráfico é acompanhada pelos vectores representativos das variáveis. A situação da Europa é determinada pelos altos valores assumidos pelas variáveis CPPNB e ENRG enquanto a situação da África é determinada pelo HMED. As restantes variáveis que não influenciam a localização visto que à escala geral não são visíveis, encontram-se na figura 5.5 como curiosidade àcerca da direcção que as orienta.

A factorização 5.12 permite, como foi referido, aproximar o comprimento de cada vector-variável ao desvio-padrão e o coseno do ângulo entre vectores à correlação entre variáveis. A figura 5.8 ilustra o "*biplot*" do Exemplo usando a factorização 5.12. Neste caso a qualidade do ajustamento da matriz aproximada é traduzida por $G_S = 0,89$. A figura 5.8 revela que a variável HMED é a que possui maior variância, logo seguida de CPPNB e de ENRG e as outras três variáveis têm variância praticamente nula. Na figura 5.9 encontra-se o "*h-plot*" do Exemplo a partir da factorização (5.12). Fazem-se duas associações de variáveis alta e positivamente correlacionadas CPPNB e ENRG, por um lado, e PURB

FIGURA 5.5 – "*BIPLOT*" para 30 paises de Africa e Europa
– Sector 1 –

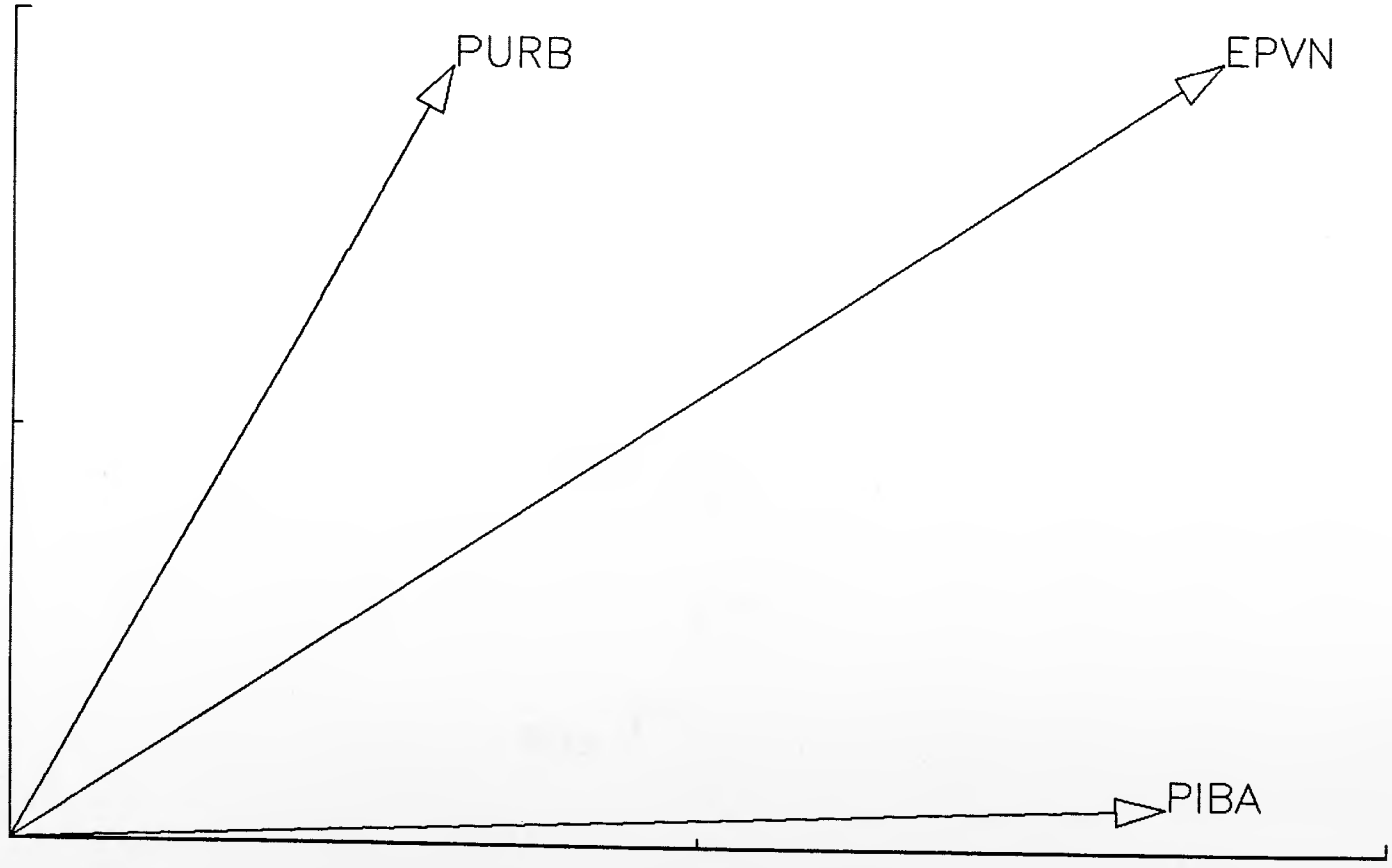


FIGURA 5.6 – "*BIPLOT*" para 30 paises de Africa e Europa
– Sector 2 –

FIGURA 5.7 – "*BIPLOT*" para 30 paises de Africa e Europa
– Sector 3 –

FIGURA 5.8 – "*BIPLOT*" para 30 paises de Africa e Europa com factorizacao (5.12)

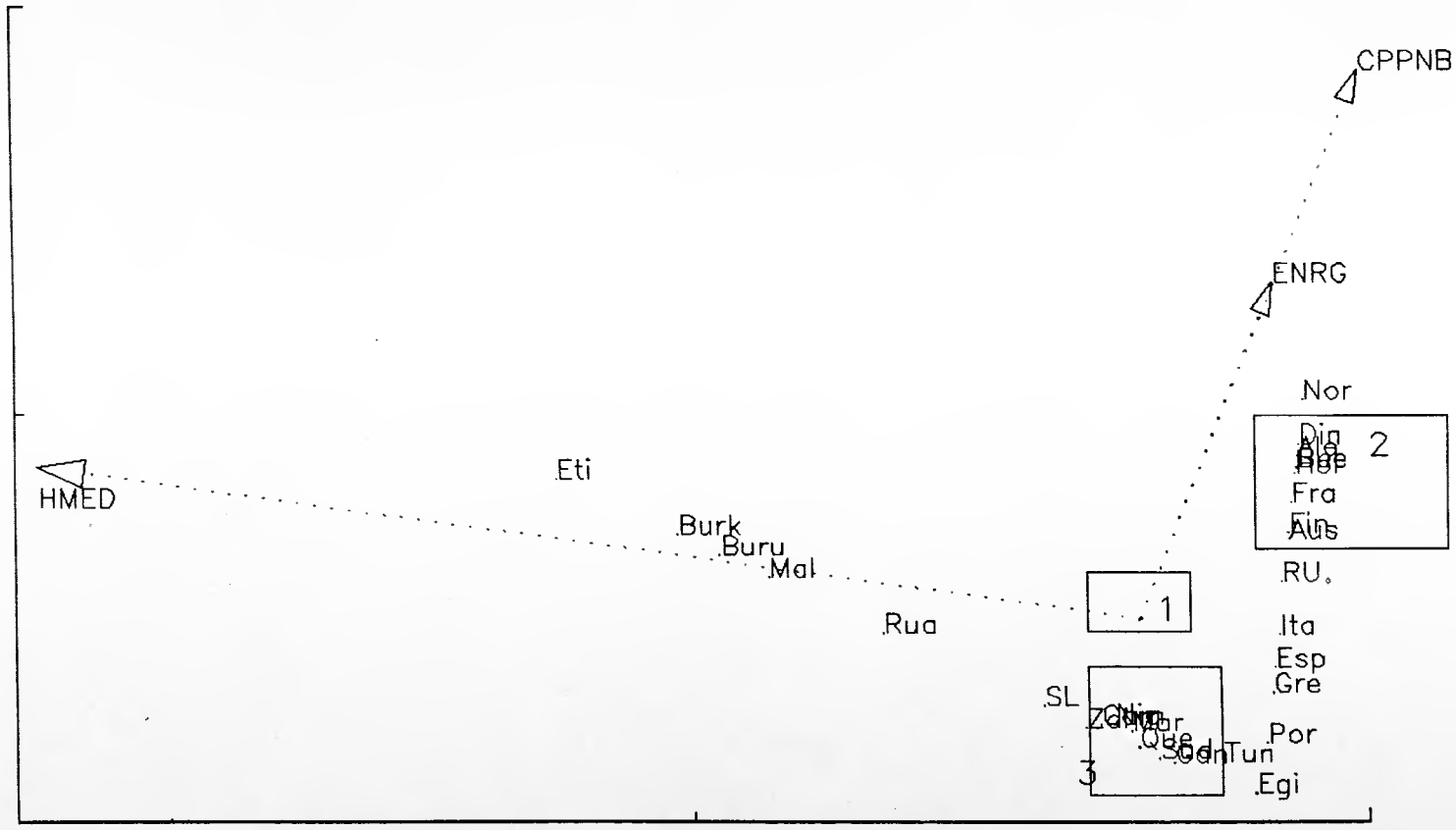
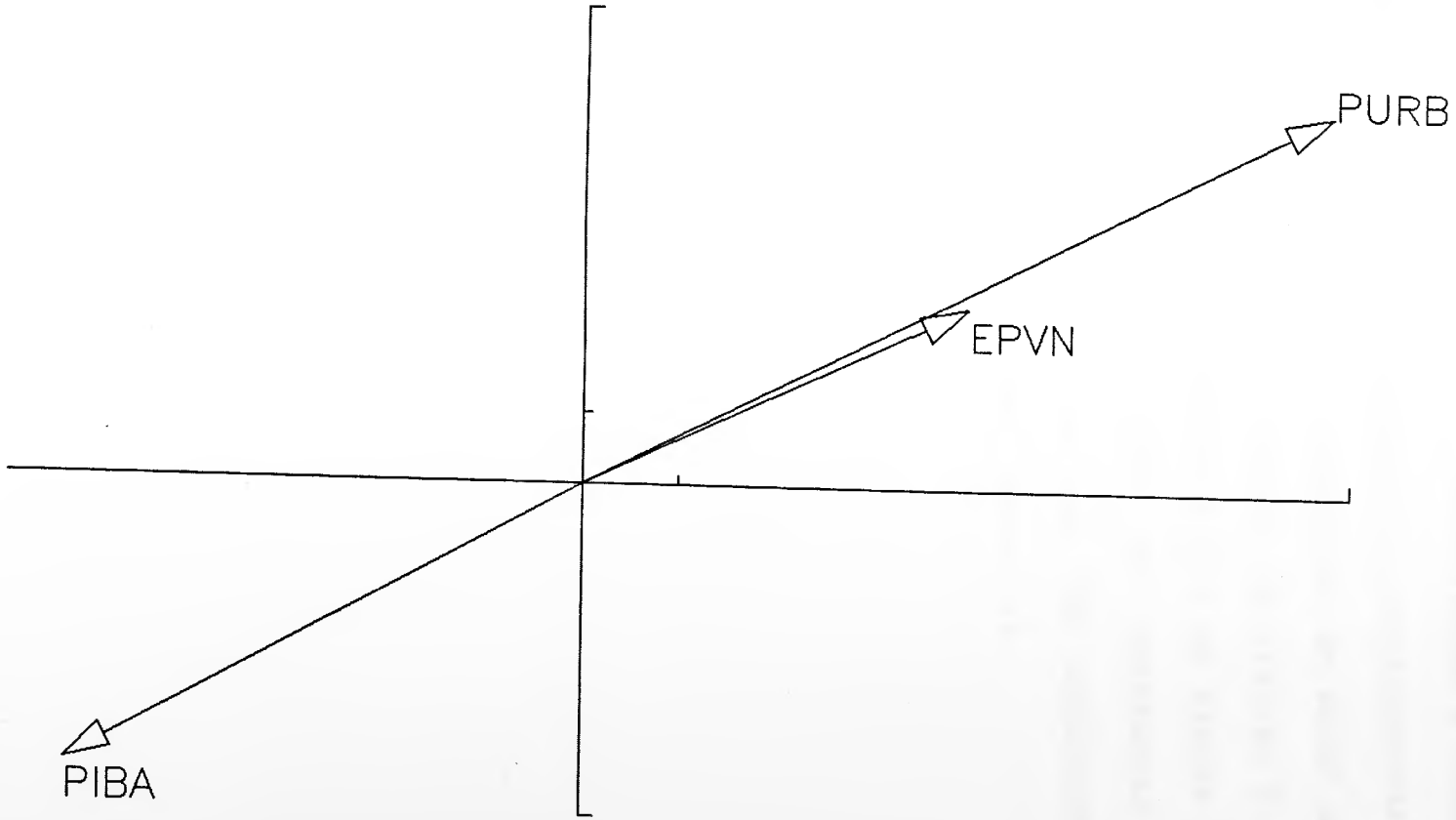


FIGURA 5.9 – "*BIPLOT*" para 30 paises de Africa e Europa com factorizacao (5.1 2) – Sector 1 –



e EPVN, por outro, verificando-se que a correlação entre CPPNB e PURB, CPPNB e EPVN, ENRG e PURB e ENRG e EPVN também é grande e em sentido directo. A variável PIBA está muito correlacionada com PURB e EPVN mas em sentido inverso, e apresenta, em menor grau, uma boa correlação negativa com CPPNB e ENRG. As figuras 5.10 e 5.11 constituem as ampliações do sectores 2 e 3 da figura 5.8. Nestas figuras encontram-se retratadas as distâncias de Mahalanobis entre os diferentes países que não modificam as associações de países feitas a partir da figura 5.4.

FIGURA 5.10 – "*BIPLOT*" para 30 paises de Africa e Europa com factorizacao (5.1 2) – Sector 2 –

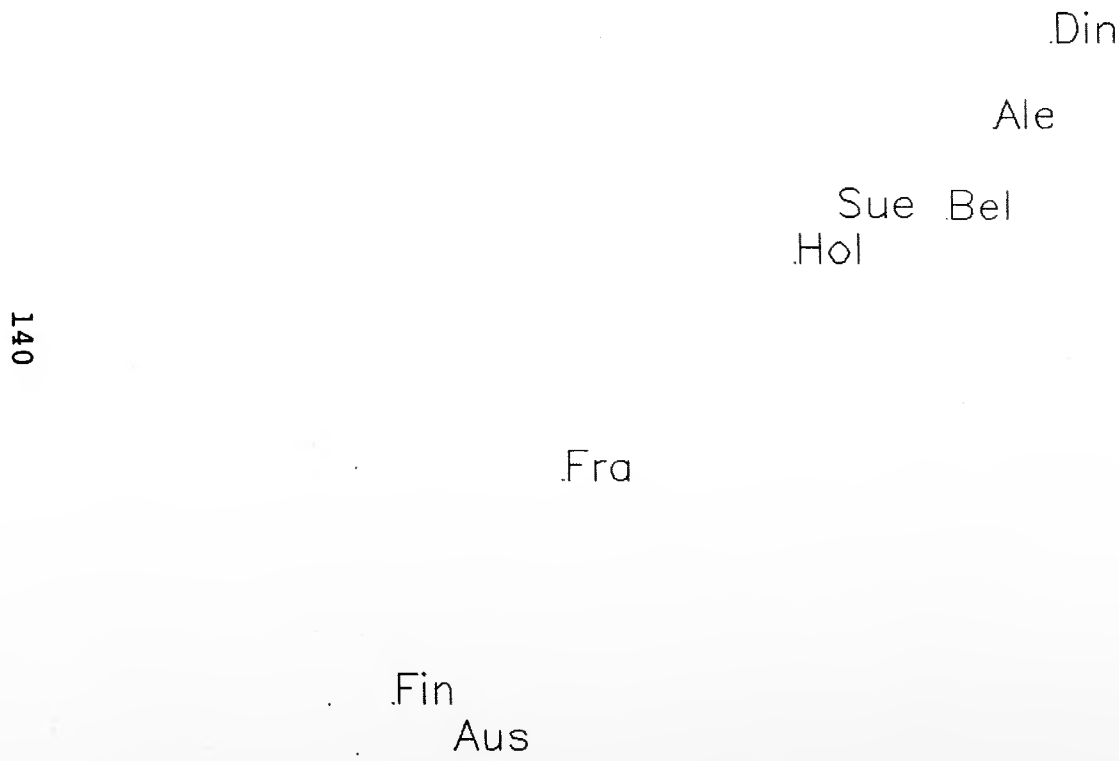
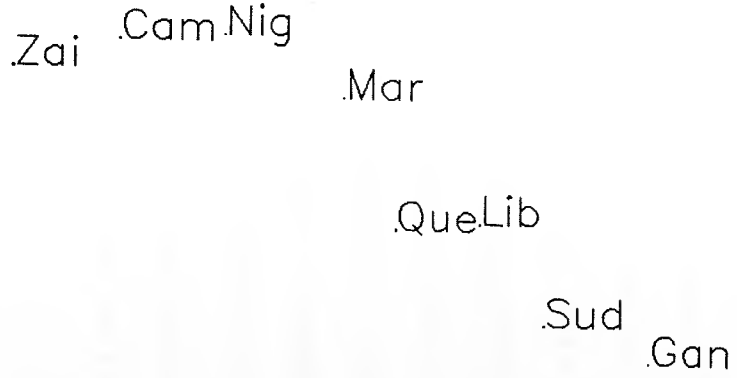


FIGURA 5.11 – "*BIPLOT*" para 30 paises de Africa e Europa com factorizacao (5.12) – Sector 3 –



CAPÍTULO 6

ESQUEMA DE REPRESENTAÇÃO TRILATERAL: "*MULTIDIMENSIONAL SCALING*" DE DIFERENÇAS INDIVIDUAIS

Os modelos aqui apresentados têm as suas origens no trabalho psicométrico, no âmbito do qual a teoria do "*scaling*" tem grande aplicabilidade. O método foi pensado para a situação em que diferentes indivíduos avaliam a dissemelhança entre pares de um conjunto de estímulos, sendo o mesmo conjunto apresentado a todos os indivíduos. Revela-se interessante examinar, não só as relações entre os estímulos, mas também, determinar que diferenças existem, se existirem, entre os conjuntos de tais relações, ou seja, investigar as diferenças que existem no conjunto das percepções individuais dos estímulos.

Os métodos de "*scaling*" expostos no capítulo 4. limitavam-se a analisar uma única matriz de dados. Se o investigador tiver à sua disposição várias matrizes, talvez uma por indivíduo, determina a matriz correspondente à média de todas as matrizes. Porém, a matriz média tem significado se se colocar a hipótese de as matrizes representarem réplicas uma das outras, ou, dito de outra

forma, as diferenças entre indivíduos se deverem a factores aleatórios. Esta hipótese é, na maioria das vezes, irrealística. Por exemplo, se a experiência consistir em questionar eleitores para que julguem a semelhança entre candidatos políticos, é muito provável que a estrutura política de um indivíduo seja muito diferente da de outro, não fazendo qualquer sentido calcular a média das opiniões dos indivíduos. Em alternativa, o investigador pode analisar cada matriz separadamente mas depara com a tarefa difícil de interpretar muitos resultados. O "*Multidimensional Scaling*" de diferenças individuais(MDSD) é o método apropriado para analisar, simultaneamente e com parsimónia, mais do que uma matriz de dados sem ter a necessidade de calcular a matriz das médias. O problema foi colocado em termos de indivíduos mas, por vezes, importa salientar diferenças entre um conjunto de ocasiões, períodos de tempo, condições experimentais, localizações geográficas, etc.. Assim, as várias matrizes em estudo, referem-se a dissemelhanças entre estímulos para diversas situações.

O primeiro procedimento de MDSD foi proposto por Tucker e Messick(1963) que o designaram "Análise de Pontos de Vista". Neste procedimento, as semelhanças julgadas por um indivíduo estão correlacionadas com as de outro indivíduo, formando uma matriz de correlações. A esta matriz de correlações é efectuada uma análise factorial: o investigador procura identificar agrupamentos de sujeitos de modo que aqueles que se posicionam

dentro dos agrupamentos estejam altamente correlacionados, ou seja, tenham julgamentos parecidos e sujeitos que se posicionam em agrupamentos diferentes tenham fracas correlações, isto é, padrões de julgamento diferentes. Após terem sido identificados os agrupamentos de indivíduos, o procedimento gera conjuntos de julgamentos hipotéticos, um conjunto para cada agrupamento. Cada conjunto de julgamentos hipotéticos é chamado um 'ponto de vista' pelos autores e pode ser interpretado como uma média dos julgamentos feitos pelos sujeitos no agrupamento correspondente; na realidade é uma média ponderada de todos os indivíduos na qual os indivíduos que pertencem ao agrupamento têm a maior ponderação. O investigador analisa cada 'ponto de vista' por um dos métodos de MDS descritos no capítulo 4. obtendo um mapa de estímulos separado e independente para cada ponto de vista.

Este procedimento não se revelou muito útil na prática e foi criticado por vários autores, nomeadamente, por Carroll e Chang(1970). Talvez a maior crítica resida no facto de não representar as diferenças individuais com parsimónia uma vez que efectua um "*scaling*" separado para cada ponto de vista. Esta abordagem é na realidade uma análise factorial seguida de "*scaling*" e não MDSD no sentido em que foi definido.

McGee(1968) propôs um procedimento não métrico cujo aspecto principal é permitir ao utilizador escolher entre uma transformação monótona separada para cada matriz ou uma única

transformação monótona para todas as matrizes. No segundo caso resulta uma só configuração do conjunto de dados pelo que não existem diferenças no estilo de julgamento dos indivíduos enquanto no primeiro se chega a um mapa por indivíduo, ou seja, processo perceptual dos vários indivíduos é tão diferente que não tem qualquer relação.

Este procedimento dispõe de um mecanismo para controlar o grau de relação entre os mapas para cada indivíduo sendo um dos extremos a não existência de relações e o outro a identidade. O mecanismo está sob o controlo do investigador não conseguindo o método determinar o melhor grau de relacionamento. Quando se assume que o estilo de julgamento é idêntico para todos os indivíduos este procedimento é equivalente a ponderar todas as matrizes e depois efectuar um só "*scaling*" não métrico. Quando se assume que o estilo de resposta difere de sujeito para sujeito, então, o procedimento é equivalente a analisar cada matriz de dados, por sua vez, por um "*scaling*" não métrico.

O método de McGee não teve grande aceitação porque, cedo, Carroll e Chang(1970) apresentaram outro, mais poderoso, assim como um programa de computador para a sua implementação ao qual foi dado o nome de "*Individual Differences Scaling*" (INDSCAL).

Antes de descrever o método importa esclarecer os conceitos fundamentais de 'espaço de grupo', 'espaço privado do sujeito *i*'

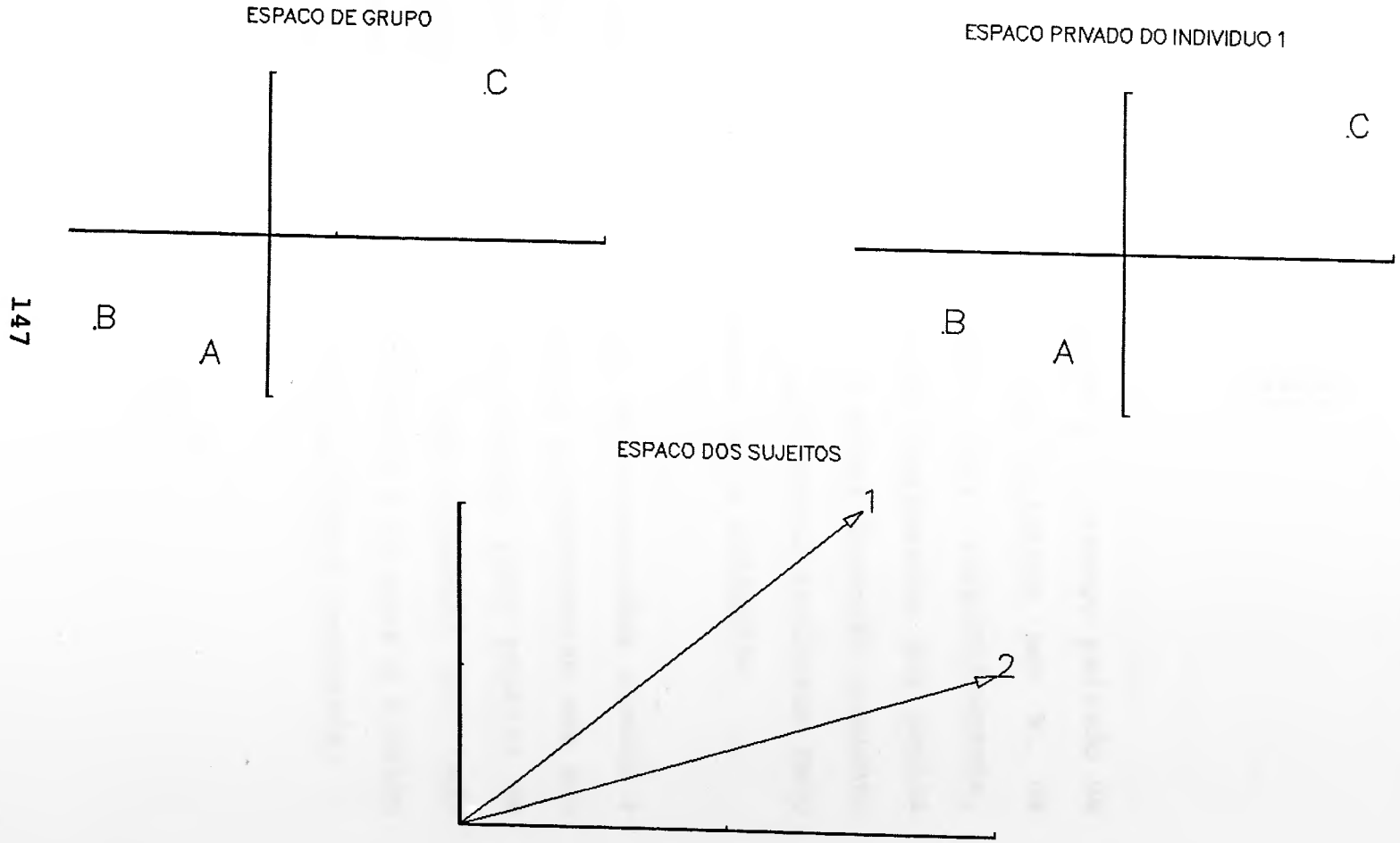
e 'espaço dos sujeitos' e suas interrelações. O espaço de grupo ou espaço dos estímulos é a configuração dos p pontos que representam os estímulos num número de dimensões t escolhido pelo investigador. O espaço privado do sujeito i é a configuração dos p pontos nas t dimensões segundo a classificação do i-ésimo indivíduo. O espaço dos sujeitos ou espaço das ponderações é, simplesmente, um modo gráfico de comparar sujeitos em termos dos seus conjuntos de ponderações e que tem, tal como o espaço de grupo t dimensões sendo cada sujeito representado por um vector localizado pelo valor das ponderações em cada uma das dimensões. Os diferentes sujeitos atribuem ponderações variáveis a cada dimensão mediante o grau de importância que cada dimensão ocupa no escala de julgamento do sujeito.

O espaço de grupo actua como uma configuração de referência para os espaços privados a partir da qual estes podem ser derivados. O espaço de grupo não descreve de facto nenhum sujeito mas revela um compromisso entre as configurações individuais com padrões de ponderações muito diferentes.

Considere-se que se dispõe das dissemelhanças entre três estímulos julgadas por dois indivíduos. Aplicou-se o modelo INSDCAL e obtiveram-se os resultados que estão representados na figura 6.1:

i) Coordenadas dos estímulos, $\mathbf{X} = \begin{bmatrix} -0.5 & -1.6 \\ -1.3 & -1.2 \\ 1.6 & 1.8 \end{bmatrix}$,

FIGURA 6.1 – MDSD para o exemplo artificial



ii) Ponderações individuais, $\mathbf{W} = \begin{bmatrix} 0.6 & 0.4 \\ 0.8 & 0.2 \end{bmatrix}$,

iii) Coordenadas dos estímulos no espaço privado do indivíduo 1,

$$\mathbf{Y}_1 = \begin{bmatrix} -0.39 & -1.01 \\ -1.01 & -0.76 \\ 1.24 & 1.14 \end{bmatrix} .$$

Designe-se, então, o espaço de grupo por **X**, o espaço privado de i-ésimo sujeito por $\mathbf{Y}_i$ e o espaço dos sujeitos por **W**. Os elementos genéricos de **X** e de $\mathbf{Y}_i$ (i=1,...,n), respectivamente, x_{jk} e y_{ijk} (j=1,...,p; k=1,...,t), são as coordenadas dos pontos que representam o j-ésimo estímulo na k-ésima dimensão enquanto w_{ik}, elemento genérico de **W**, reflete a importância atribuída pelo indivíduo i à dimensão k na sua percepção dos p estímulos.

O modelo assume que existe um conjunto de t dimensões comuns a todos os sujeitos nas quais os p estímulos se representam mas que as distâncias entre pontos diferem de sujeito para sujeito de acordo com a ponderação atribuída a cada dimensão por cada sujeito. A dissemelhança entre os estímulos j e r, para o i-ésimo indivíduo é bem traduzida pela distância euclideana ponderada,

$$(6.1) \qquad d^2_{ijr} = \sum_{k=1}^{t} w_{ik}(x_{jk} - x_{rk})^2 .$$

Os espaços privados podem ser derivados do espaço de grupo através de uma simples transformação,

$$y_{ijk} = x_{jk} \sqrt{w_{ik}} \quad , \tag{6.2}$$

ou seja, as coordenadas dos espaços privados de cada sujeito são simplesmente uma versão ponderada das coordenadas do espaço de grupo obtidas por rescalonamento (contracção ou expansão) dos eixos do espaço de grupo.

As ponderações w_{ik} podem assumir quaisquer valores reais não negativos. Se $w_{ik}=0$ significa que a k-ésima dimensão não entra no espaço privado i. Se $w_{ik}=1$, então a k-ésima dimensão é exactamente a mesma do espaço de grupo. Se $w_{ik}<1$ o i-ésimo sujeito contrai a k-ésima dimensão e se $w_{ik}>1$ o i-ésimo sujeito expande a k-ésima dimensão relativamente ao espaço de grupo.

Considerem-se as n matrizes de dissemelhança Δ_i cujo elemento genérico δ_{ijr} é o valor da dissemelhança entre os j e r-ésimos estímulos para o i-ésimo indivíduo. O modelo que aqui se descreve é visto, inicialmente, como um problema de MDSC para cada matriz Δ_i. Assim, para obter as coordenadas y_{ijk} do espaço privado $\mathbf{Y}_i$ começa por definir-se a matriz,

$$\mathbf{Q}_i = \mathbf{Y}_i \, \mathbf{Y}_i^T \quad , \tag{6.3}$$

cujos elementos q_{ijr} se calculam,

$$q_{ijr} = -\frac{1}{2}\left[\delta^2_{ijr} - \delta^2_{ij.} - \delta^2_{i.r} + \delta^2_{i..}\right]. \tag{6.4}$$

(Compare-se esta expressão com (4.6)). A relação (6.2) escreve-se matricialmente,

$$\mathbf{Y}_i = \mathbf{X}\,\mathbf{W}_i^{1/2}, \tag{6.5}$$

onde $\mathbf{W}_i = \text{diag}(w_{i1},\ldots,w_{it})$ é a matriz diagonal que contém as ponderações atribuídas pelo i-ésimo indivíduo às t dimensões. Neste caso,

$$\mathbf{Q}_i = \mathbf{X}\,\mathbf{W}_i\,\mathbf{X}^T, \tag{6.6}$$

cujo elemento genérico se expressa,

$$q_{ijr} = \sum_{k=1}^{t} w_{ik}\, x^L_{jk}\, x^R_{rk}. \tag{6.7}$$

Os índices L e R servem apenas para distinguir a matriz **X** que se encontra à esquerda da matriz **X** que se encontra à direita. O conjunto dos valores das matrizes **X** e **W** encontra-se utilizando iterativamente o princípio dos mínimos quadrados através de um

procedimento denominado "Decomposição canónica de tabelas de n-entradas" da autoria de Eckart e Young (1936). Carroll e Chang(1970) aplicaram-no no INDSCAL da seguinte forma:

1) Atribuem-se estimativas iniciais para $\mathbf{X}^L$ e $\mathbf{X}^R$ (podem ser e são geralmente iguais) e procura-se uma estimativa de mínimos quadrados para $\mathbf{W}$.

2) Seja $u=p(j-1)+r$ $(u=1,\ldots,p^2)$ e defina-se,

$$g_{uk} = x^L_{jk}\, x^R_{rk} \ ,$$

e,

$$q^*_{iu} = q_{ijk}.$$

Então (6.7) pode escrever-se,

$$q^*_{iu} \cong \sum_{k=1}^{t} w_{ik}\, g_{uk} \ , \tag{6.8}$$

em que o símbolo $\cong$ significa que se procura uma solução de mínimos quadrados para os parâmetros da direita. Esta equação escrita na forma matricial conduz a,

$$\mathbf{Q}^* \cong \mathbf{W}\,\mathbf{G}^T \ , \tag{6.9}$$

onde $\mathbf{Q}^*(n\times p^2)$ tem elemento genérico q^*_{iu} e $\mathbf{G}(p^2\times t)$ de elemento g_{uk}

respeita a igualdade $\mathbf{G}=\text{vec}(\mathbf{X}^L)\otimes\text{vec}(\mathbf{X}^R)$. A solução de mínimos quadrados para $\mathbf{W}$ é,

$$\hat{\mathbf{W}} = \mathbf{Q}^* \mathbf{G} (\mathbf{G}^T\mathbf{G})^{-1}. \tag{6.10}$$

3) Obtida uma estimativa para $\mathbf{W}$ procura obter-se uma melhor estimativa para $\mathbf{X}^L$ de modo semelhante. Seja $v = n(i-1)+r$ $(v=1,\ldots,np)$ e defina-se,

$$h_{vk} \equiv w_{ik}\, x^R_{rk} ,$$

e,

$$q^{**}_{rv} \equiv q_{ijr}.$$

A equação (6.7) pode ser escrita,

$$q^{**}_{rv} \cong \sum_{k=1}^{t} x^L_{jk}\, h_{vk} , \tag{6.11}$$

ou,

$$\mathbf{Q}^{**} \cong \mathbf{X}^L \mathbf{H}^T , \tag{6.12}$$

onde $\mathbf{Q}^{**}(p\text{x}np)$ tem elemento genérico q^{**}_{rv} e $\mathbf{H}(np\text{x}p)$ de elemento genérico h_{vk} é $\mathbf{H}=\text{vec}(\mathbf{W})\otimes\text{vec}(\mathbf{X}^R)$. A solução de mínimos quadrados para $\mathbf{X}^L$ é,

$$\hat{\mathbf{X}}^L = \mathbf{Q}^{**} \mathbf{H} (\mathbf{H}^T\mathbf{H})^{-1} . \tag{6.13}$$

4) Dados os novos valores para $\mathbf{W}$ e $\mathbf{X}^L$ procura-se uma melhor estimativa para $\mathbf{X}^R$ pelo mesmo processo.

O ciclo é repetido (usam-se os valores do conjunto de parâmetros obtidos na iteração anterior para melhorar as estimativas dos parâmetros) até que um ciclo completo seja efectuado sem alteração dos parâmetros para um dado nível de precisão previamente definido. Repare-se que em cada passo deste processo iterativo se está a reduzir ou, pelo menos, a não aumentar o total da soma dos erros quadráticos,

$$\sum_i \sum_j \sum_k \varepsilon_{ijk}^2 \,, \tag{6.14}$$

originados pelo ajustamento,

$$q_{ijr} = \sum_{k=1}^{t} \hat{w}_{ik}\, \hat{x}^L_{jk}\, \hat{x}^R_{rk} + \varepsilon_{ijk} \,, \tag{6.15}$$

visto que a solução de mínimos quadrados é precisamente a que minimiza a soma (6.14).

Se o algoritmo descrito conduz a diferentes estimativas para as matrizes $\mathbf{X}^L$ e $\mathbf{X}^R$ acrescenta-se um ciclo final que consiste em igualar as duas matrizes tomando a estimativa da que foi obtida

em último lugar e recalcular **W**.

Uma questão importante neste método é a normalização a que são submetidos quer os dados iniciais quer a solução final. A normalização dos dados iniciais consiste em centrar cada matriz de dados, subtraindo aos respectivos elementos a média da coluna. A centralização das matrizes localiza o centróide do espaço de grupo na origem e implica atribuir a cada sujeito igual influência na solução do modelo.

Fenómeno novo no INDSCAL é a normalização das ponderações w_{ik}. A necessidade desta normalização surge porque as distâncias d^2_{ijr} expressas por (6.1) não se alteram quando, simultaneamente, cada dimensão é contraída ou expandida por um factor de escala a_k e w_{ik} é multiplicado pela constante $(1/a_k^2)$,

$$d^2_{ijr} = \sum_{k=1}^{t} w_{ik} \, (1/a_k^2) \, (\, a_k x_{jk} - a_k x_{rk})^2 \, . \tag{6.16}$$

A normalização das ponderações, proposta pelos autores, é,

$$\sum_{k=1}^{t} w_{ik}^2 = R_i^2 \, , \tag{6.17}$$

onde R_i^2 significa a proporção da variância das coordenadas do

sujeito *i* explicada pelo modelo.

As dimensões resultantes do INDSCAL são unicamente identificadas, isto é, se as dimensões sofrerem uma rotação e as ponderações recalculadas, a solução obtida explica os dados menos bem que a solução original. Apesar do espaço dos estímulos não poder sofrer rotações é possível efectuar-lhe reflexões e translações preservando toda a informação significativa.

Escolheram-se os países europeus que integram o Quadro 2.1 e recolheu-se uma nova amostra para os mesmos indicadores no ano de 1990. Às duas matrizes de dados aplicou-se o modelo INDSCAL cujo ajustamento foi feito em oito iterações, tendo sido obtido um bom valor para a soma dos resíduos quadráticos de 0,971.

O resultado do modelo em termos dos espaço de grupo, espaço dos sujeitos e espaços privados está expresso nas figuras 6.2 a 6.5. O espaço de grupo deve ser interpretado com cautela; na realidade representa um ano fictício que pondera igualmente as duas dimensões. Se a relação de ponderações para cada ano difere de um é um perigo tentar interpretar uma configuração que não é representativa.

No espaço dos sujeitos o vector que traduz a igualdade de ponderações é desenhado com um ângulo de 45° e vectores situados na mesma direcção têm o mesmo quociente de ponderações. A partir

FIGURA 6.2 – MDSD para 14 paises da Europa, para os anos de 1980 e 1990 (ESPACO DE GRUPO)

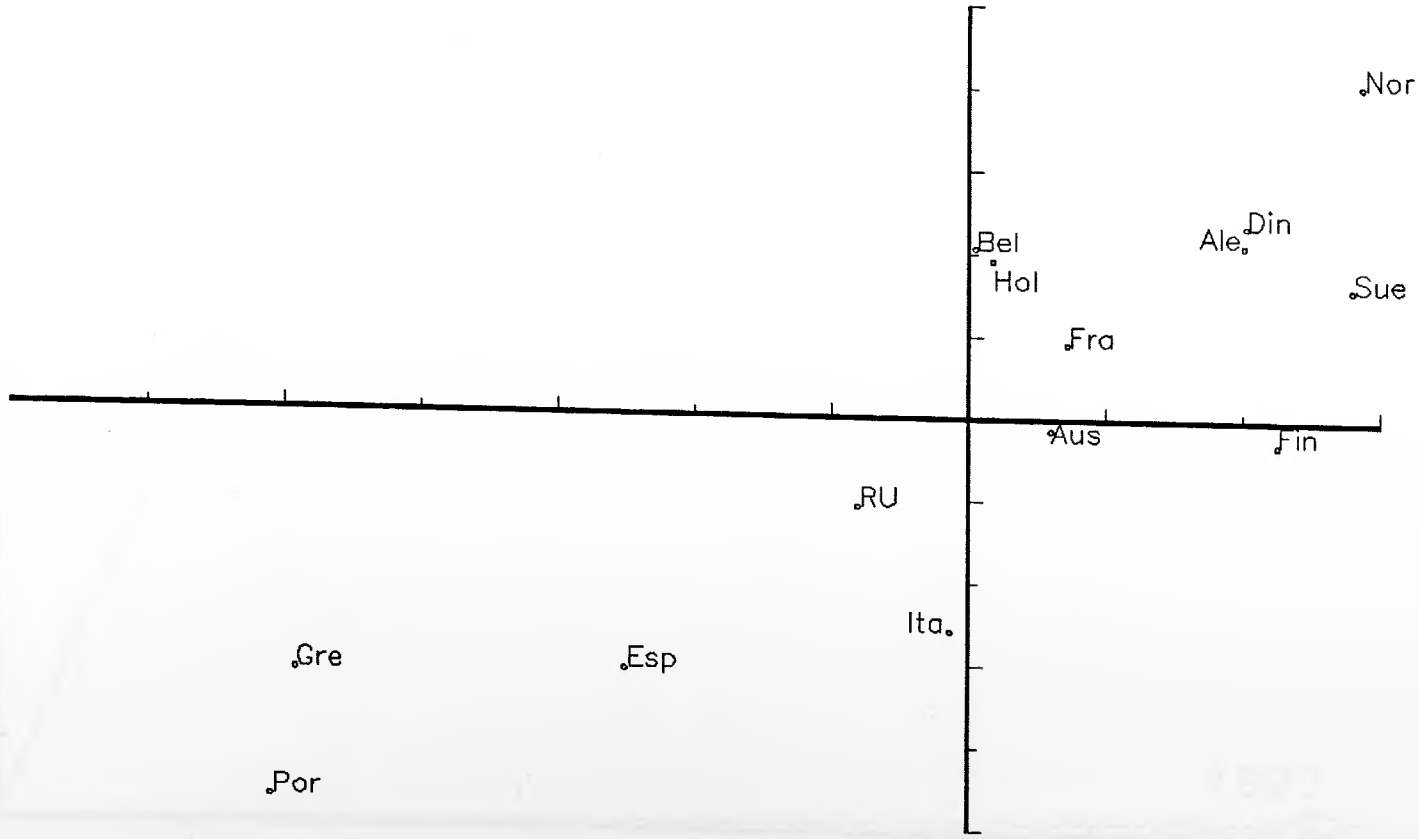


FIGURA 6.3 – MDSD para 14 paises da Europa, para os anos de 1980 e 1990 (ESPACO DOS SUJEITOS)

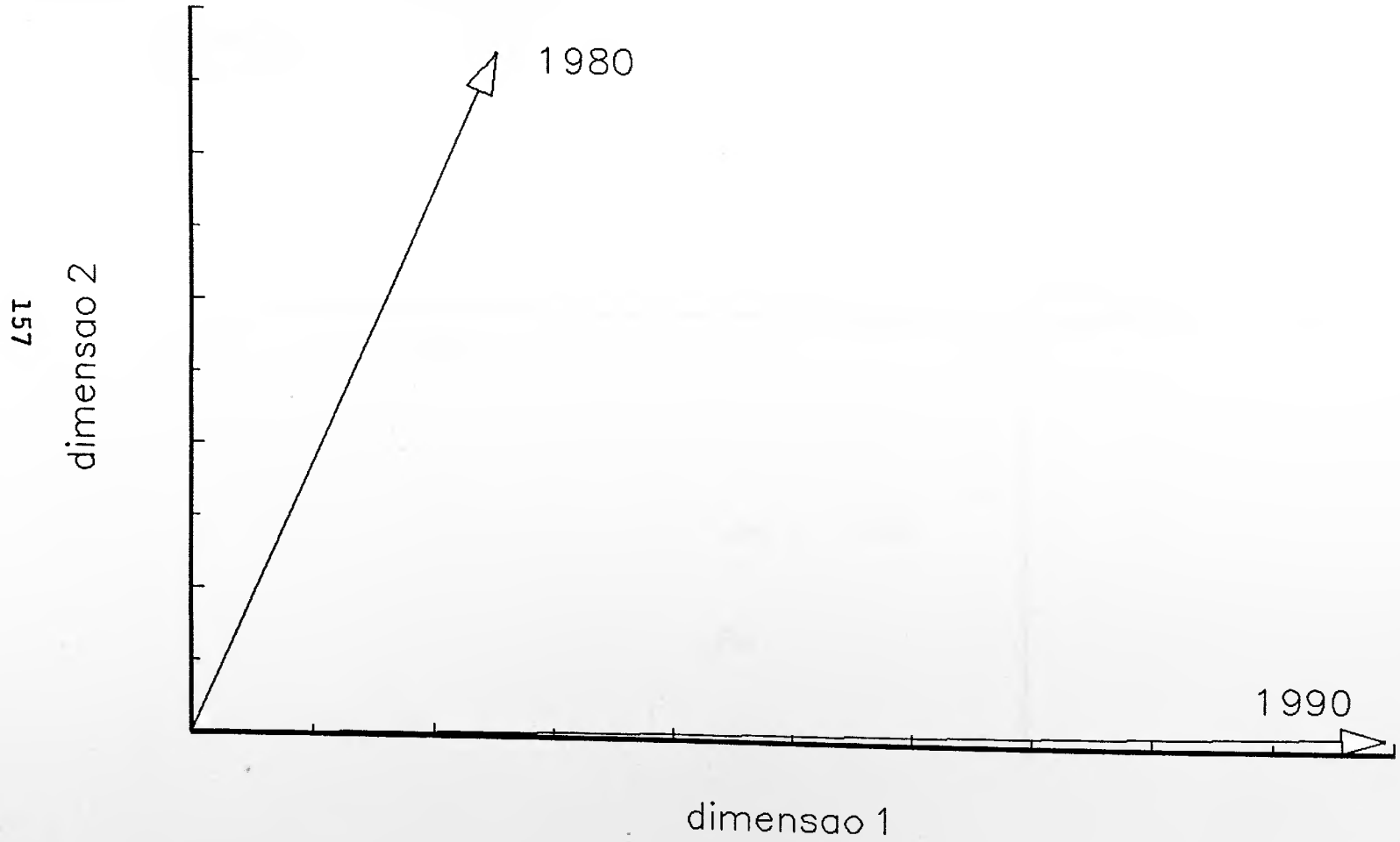


FIGURA 6.4 – MDSD para 14 paises da Europa, para os anos de 1980 e 1990 (ESPACOS PRIVADOS)

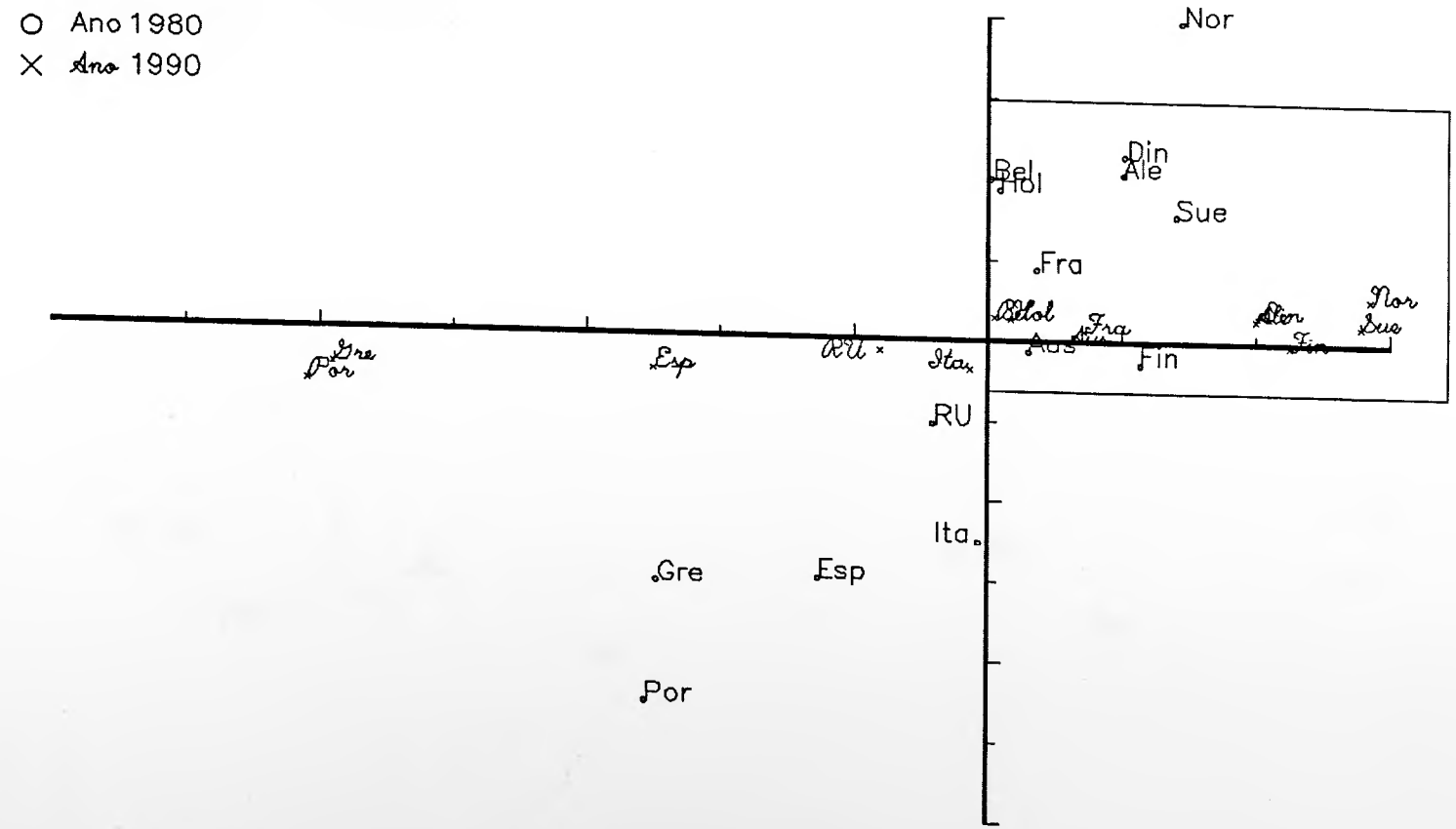
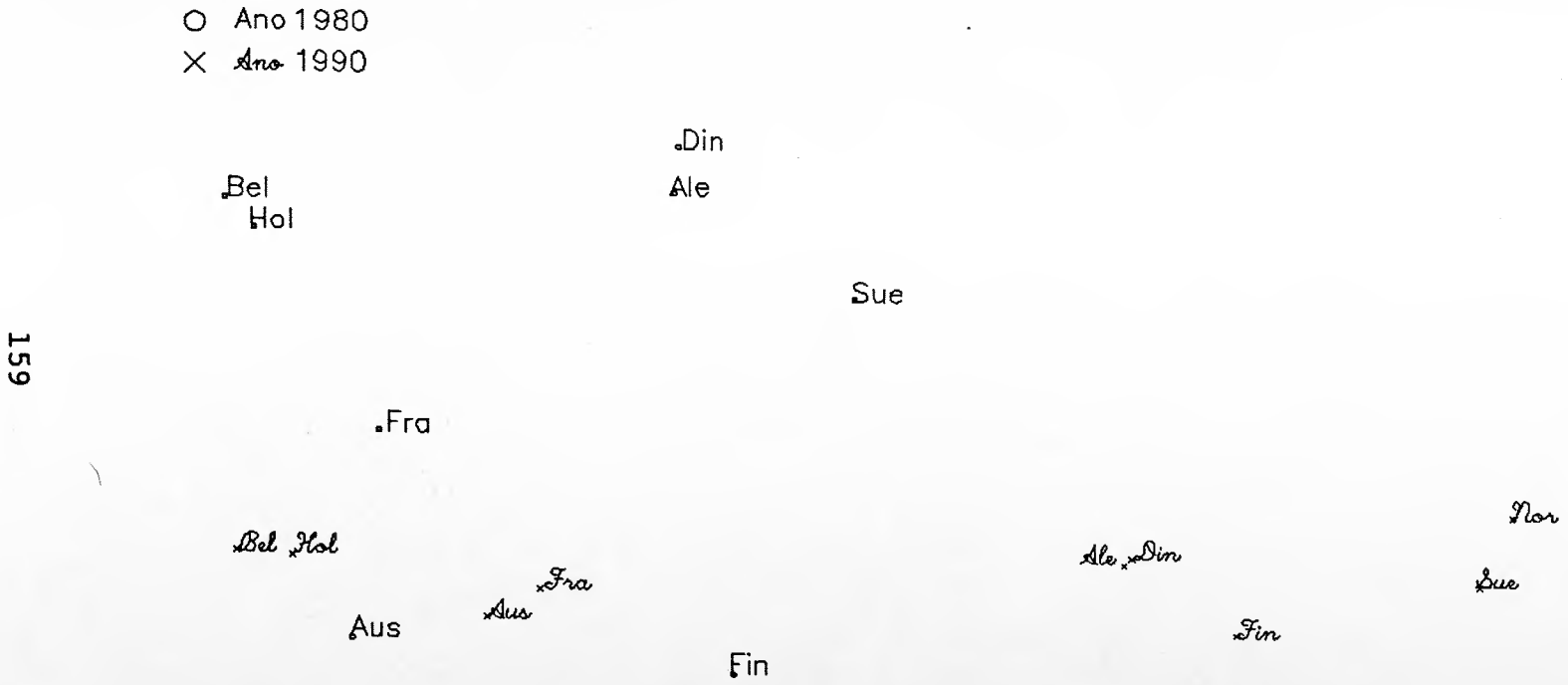


FIGURA 6.5 – MDSD para 14 paises da Europa, para os anos de 1980 e 1990 (ESPACOS PRIVADOS) – Sector 1



da figura 6.3 verifica-se que as duas dimensões não são igualmente ponderadas por nenhum dos anos e, ainda, que os dois anos têm padrões de ponderações muito diferentes. Com efeito, o ano de 1980 dá preponderância à dimensão 2 enquanto o ano de 1990 atribui maior importância à dimensão 1. A normalização (6.17) permite avaliar a adequação de cada espaço individual aos dados individuais através do comprimento dos vectores. Se este comprimento é próximo de um, o que acontece para ambos os anos, então o ajustamento é muito bom.

O espaço de grupo revela a sua verdadeira função ao viabilizar a determinação dos espaços privados de cada ano que aparecem sobrepostos no gráfico da figura 6.4. A sobreposição dos espaços privados permite observar a evolução das posições relativas entre os países nos dez anos em estudo. A Grécia atrasou-se face aos restantes países europeus, juntando-se a Portugal numa situação extrema. A Itália melhorou a situação económico-social revelado pelo percurso em direcção aos países do Norte, ultrapassando mesmo o Reino Unido. No grupo dos países da frente salienta-se o menor afastamento da Noruega, da Suécia e da Finlândia que alinharam com a Alemanha e a Dinamarca. O sector assinalado encontra-se ampliado na figura 6.5.

Ensaiou-se uma análise a três dimensões que originou um valor mais reduzido para a soma dos resíduos quadráticos, da ordem de 0,994, mas a melhoria não justifica o uso de mais uma dimensão.

CAPÍTULO 7

COMPARAÇÃO DE REPRESENTAÇÕES: ANÁLISE PROCRUSTEANA

Acontece frequentemente, no decurso de uma investigação, serem utilizados, para os mesmos dados, diferentes métodos de representação gráfica ou diferentes medidas de dissemelhança ou, ainda, proceder-se à recolha de uma nova matriz de dados. Em qualquer destes casos são obtidas várias configurações sendo desejável compará-las. Quando se obtém uma nova matriz de dados pode usar-se, em alternativa, o método INSDCAL (vidé capítulo 6).

A análise procrusteana é uma técnica que mede a divergência entre as várias representações. A designação de procrusteana deriva da semelhança com as actividades do criminoso mitológico Procrustes. Este assegurava que os trabalhadores se ajustavam às camas que lhes eram destinadas esticando-os ou cortando-lhes as pernas.

Originalmente, a análise procrusteana foi desenvolvida para ser usada na sequência de uma análise factorial mas autores como Schönemann e Carrol(1970), Gower(1971b) e Sibson(1978) impulsionaram a aplicação da técnica ao espírito do

"*Multidimensional Scaling*".

No capítulo 4. a compreensão dos gráficos faz-se através das distâncias relativas existentes entre os pontos. Duas configurações dizem-se iguais, ou seja, permitem a mesma leitura, quando envolvam apenas diferenças de escala, orientação e origem. As representações podem sofrer,

i) translações,
ii) rotações ortogonais e/ou reflexões,
iii) mudanças de escala,

dos eixos sem que a distância relativa entre pontos seja alterada e as características das configurações se modifiquem.

A translação de um gráfico não é mais que um deslocamento fixo de todos os seus pontos através de uma distância constante numa direcção comum, isto é, a origem é transferida. Uma rotação ortogonal consiste no deslocamento fixo de todos os pontos através de um ângulo constante mantendo inalterada a distância de cada ponto ao centróide e a ortogonalidade dos eixos. A reflexão de um eixo acontece quando os sinais de todas as coordenadas relativamente a esse eixo são trocados. Reflectir um número par de dimensões é equivalente a uma rotação ortogonal pelo que, no texto, o termo reflexão se aplica sempre a um número ímpar de dimensões. Pode mesmo descrever-se a reflexão como uma rotação

ortogonal do maior número par de dimensões contido naquele número ímpar seguida da reflexão de apenas uma dimensão. Finalmente, chama-se mudança de escala a uma expansão ou contracção de todos os pontos de uma configuração por um valor constante numa linha recta a partir do centróide.

Estas operações geométricas podem ser apresentadas em linguagem matricial. Sendo **X**, a matriz das coordenadas dos pontos da configuração, a translação corresponde à adição de uma matriz de linhas idênticas, a rotação e a reflexão à multiplicação por uma matriz ortogonal e a mudança de escala à multiplicação por uma constante.

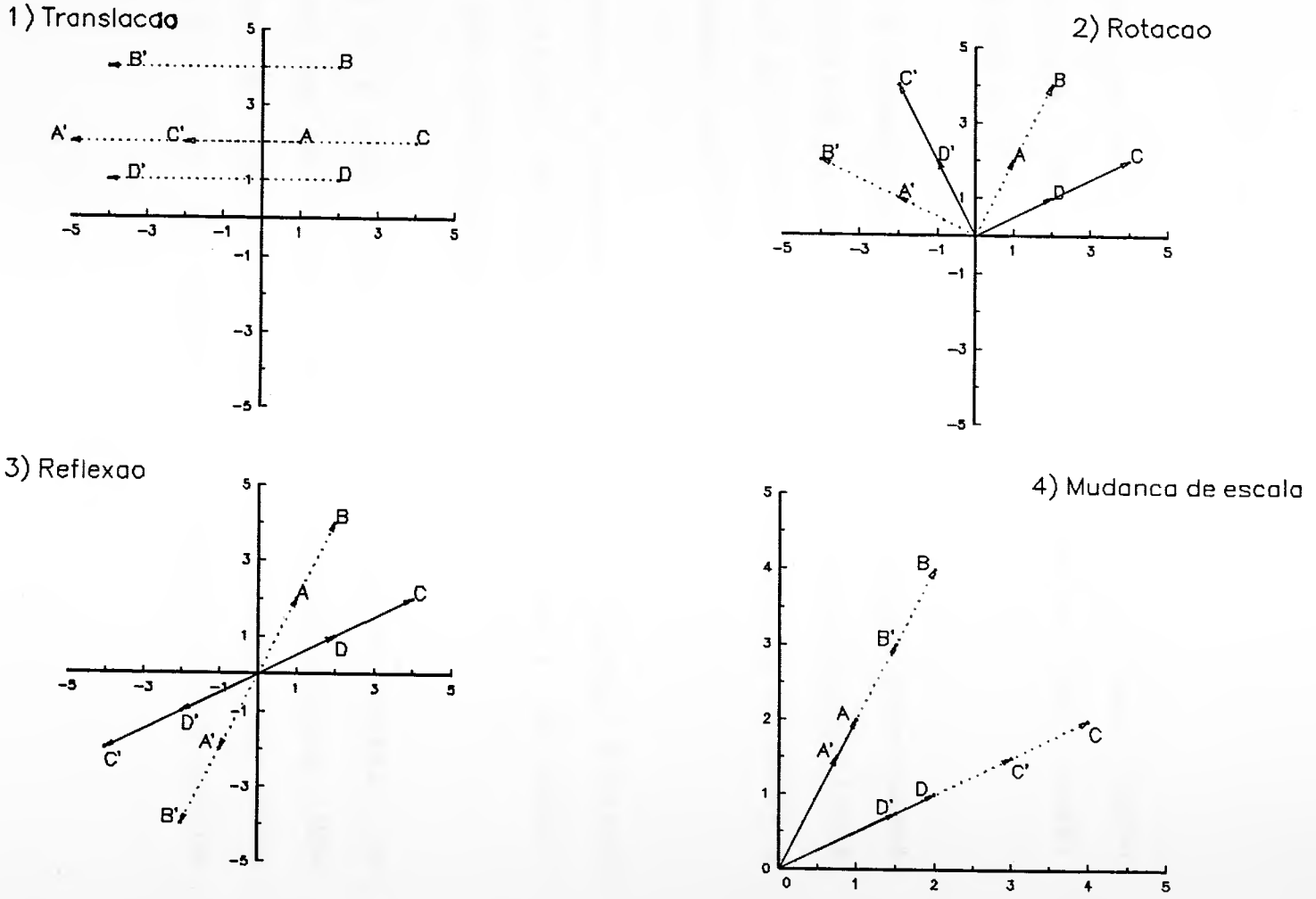
A figura 7.1 ilustra as transformações com os pontos de $\mathbb{R}^2$, A, B, C e D cuja matriz **X** de coordenadas é,

$$\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 4 & 2 \\ 2 & 1 \end{bmatrix} .$$

Em 1) efectuou-se uma translação adicionando a matriz A,

$$\mathbf{A} = \begin{bmatrix} -6 & 0 \\ -6 & 0 \\ -6 & 0 \\ -6 & 0 \end{bmatrix} .$$

FIGURA 7.1 – Transformacoes geometricas



Em 2) fez-se uma rotação de 90°, pós-multiplicando **X** pela matriz ortogonal **Q**,

$$\mathbf{Q} = \begin{bmatrix} \cos 90° & \text{sen } 90° \\ -\text{sen } 90° & \cos 90° \end{bmatrix}.$$

Em 3) efectuou-se uma reflexão, o que é equivalente a uma rotação de 180°. Finalmente, em 4), mudou-se a escala dos eixos, multiplicando **X** por 0.75.

Antes de medir a divergência entre as configurações efectuam-se as operações descritas para tornar os gráficos o mais compatíveis possível. Uma vez que tal tenha sido feito é o momento de medir a falta de ajustamento que ainda persista.

Refira-se, primeiro, a comparação de duas representações. É claro que as representações em causa têm que se referir às mesmas entidades para que faça sentido compará-las.

Considere-se **X** e **Y** duas matrizes contendo as coordenadas de *n*-pontos em cada uma de duas configurações *p*-dimensionais. Uma medida simples do grau de coincidência entre as duas representações é a soma das distâncias quadradas entre pontos homólogos,

$$(7.1) \qquad M^2 = \sum_{i=1}^{n} [\sum_{j=1}^{p} (x_{ij} - y_{ij})^2] \ .$$

Um valor nulo de M^2 significa que os gráficos são idênticos.

O procedimento adoptado mantém uma configuração fixa e efectua, sequencialmente, as operações geométricas i) a iii) à outra. Seja **X** a configuração que se mantém e **Y** aquela que sofre as transformações.

i) Ajustamento após translação

O primeiro passo consiste na centralização das duas matrizes **X** e **Y**. Os centróides das representações designam-se G_X e G_Y e têm como coordenadas, respectivamente, $\bar{\mathbf{x}}_j$ e $\bar{\mathbf{y}}_j$ que se definem,

$$(7.2) \qquad \bar{\mathbf{x}}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

e,

$$\bar{\mathbf{y}}_j = \frac{1}{n} \sum_{i=1}^{n} y_{ij} \qquad (j=1,\ldots,p) \ .$$

A medida M^2 calculada para as matrizes centradas é,

$$(7.3)^* \quad M^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} \{(x_{ij} - \bar{\mathbf{x}}_j) - (y_{ij} - \bar{\mathbf{y}}_j)\}^2 + n \sum_{j=1}^{p} (\bar{\mathbf{x}}_j - \bar{\mathbf{y}}_j)^2 .$$

Designando por M_O^2 a soma dos resíduos ao quadrado entre as duas configurações após os centróides terem sido transferidos para a origem, M^2 pode escrever-se,

$$M^2 = M_O^2 + n \sum_{j=1}^{p} (\bar{x}_j - \bar{y}_j)^2 \quad . \tag{7.4}$$

O somatório $\sum_{j=1}^{p} (\bar{x}_j - \bar{y}_j)^2$ não é mais que o valor da distância euclideana entre os dois centróides G_X e G_Y, simbolizada por $\overline{G_X G_Y}$, pelo que,

$$M^2 = M_O^2 + n \, (\overline{G_X G_Y})^2 . \tag{7.5}$$

O melhor ajustamento entre as matrizes **X** e **Y**, após a translação, ocorre quando o centróide de ambas as configurações é a origem.

ii) Ajustamento após rotação e/ou reflexão

O procedimento sequencial adoptado determina que se considerem as matrizes **X** e **Y** centradas na origem. Uma rotação de **Y** relativamente a **X** exprime-se por uma matriz ortogonal **Q** e após a rotação as coordenadas da configuração são dadas pelas linhas de **YQ**. A medida M^2 apresenta-se,

(7.6)* $$M^2 = tr(\mathbf{X}\mathbf{X}^T) + tr(\mathbf{Y}\mathbf{Y}^T) - 2\ tr(\mathbf{Q}^T\mathbf{Y}^T\mathbf{X}) \ .$$

O melhor ajustamento entre as matrizes **X** e **Y**, após a rotação, é obtido quando a matriz de rotação **Q** se calcula através de,

(7.7)* $$\mathbf{Q} = \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X})^{-1/2} \ .$$

A esta operação chama-se rotação procrusteana de **Y** relativamente a **X**. O valor mínimo de M^2 é dado por,

(7.8)* $$M^2 = tr(\mathbf{X}\mathbf{X}^T) + tr(\mathbf{Y}\mathbf{Y}^T) - 2\ tr(\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X})^{1/2}.$$

Uma vez que,

(7.9) $$tr(\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X})^{1/2} = tr(\mathbf{Y}^T\mathbf{X}\mathbf{X}^T\mathbf{Y})^{1/2} \ ,$$

os papéis de **X** e **Y** podem ser invertidos sem alterar o valor óptimo de M^2: é indiferente ajustar **X** a **Y** ou **Y** a **X**.

iii) Ajustamento após mudança de escala

Se existirem diferenças de escala entre as duas configurações o passo final deste procedimento consiste em multiplicar as coordenadas da matriz **Y** por um escalar *c* e estimar o valor de *c* que minimiza M^2. As operações são sequenciais pelo que **Y** já foi

centrada e sofreu uma rotação para $\mathbf{YQ}$. A multiplicação por c converte as coordenadas em $c\mathbf{YQ}$ e,

$$M^2 = \mathrm{tr}(\mathbf{XX}^T) + c^2\mathrm{tr}(\mathbf{YY}^T) - 2c\ \mathrm{tr}(\mathbf{Q}^T\mathbf{Y}^T\mathbf{X}). \tag{7.10}$$

A aplicação de simples cálculo diferencial determina que o mínimo valor de M^2 seja obtido com,

$$c = \frac{\mathrm{tr}(\mathbf{Q}^T\mathbf{Y}^T\mathbf{X})}{\mathrm{tr}(\mathbf{YY}^T)} . \tag{7.11}$$

Com os resultados da alínea anterior,

$$c = \frac{\mathrm{tr}(\mathbf{X}^T\mathbf{YY}^T\mathbf{X})^{1/2}}{\mathrm{tr}(\mathbf{YY}^T)} , \tag{7.12}$$

e,

$$M^2 = \mathrm{tr}(\mathbf{XX}^T) - \frac{\mathrm{tr}^2(\mathbf{X}^T\mathbf{YY}^T\mathbf{X})^{1/2}}{\mathrm{tr}(\mathbf{YY}^T)} . \tag{7.13}$$

A análise efectuada em ii), que determina a melhor rotação, é independente de qualquer factor de escala c. O cálculo do c óptimo não influencia a determinação de $\mathbf{Q}$ feita por (7.7).

Mas o processo de ajustamento de uma configuração a outra não é simétrico no caso da operação iii). Sendo a matriz $\mathbf{X}$ a sofrer os

efeitos de escala as coordenadas convertem-se em $c\mathbf{XQ}$ e M^2 em,

$$M^2 = c^2 \, tr(\mathbf{XX}^T) + tr(\mathbf{YY}^T) - 2c \, tr(\mathbf{Q}^T\mathbf{Y}^T\mathbf{X}) \quad . \tag{7.14}$$

Para evitar problemas de interpretação Gower(1971b) propôs estandardizar as matrizes **X** e **Y**, de modo que,

$$tr(\mathbf{XX}^T) = tr(\mathbf{YY}^T) = 1 \quad . \tag{7.15}$$

Agora,

$$c = tr(\mathbf{X}^T\mathbf{YY}^T\mathbf{X})^{1/2} \quad , \tag{7.16}$$

é um valor independente da escolha feita: a configuração **X** ajusta-se a **Y** ou vice-versa.

Considerou-se, ao longo do texto, que ambas as configurações são p-dimensionais. Se, porém, o número de dimensões for diferente deve aumentar-se a matriz correspondente ao gráfico com menos dimensões, com colunas de zeros até que ambas as matrizes possuam o mesmo número de colunas e, deste modo, realizar a análise.

Para comparar mais do que duas representações pode efectuar-se o procedimento anterior para todas as combinações de pares possíveis de representações e analisar a proximidade das várias medidas M^2 através de um método de "*scaling*" .

Gower(1975) alargou a ideia da análise procrusteana por forma a que numa única análise várias configurações sofram simultaneamente rotações, translações, reflexões e mudanças de escala e, apenas, um critério de ajustamento seja optimizado.

Seja $\mathbf{X}^{(i)}$, $(i=1,\ldots g)$, a matriz que contém as coordenadas de n pontos em p dimensões para a i-ésima configuração. Assume-se que a s-ésima coluna de todas as matrizes $\mathbf{X}$ se refere à mesma entidade para que as g configurações sejam comparáveis. Assume-se, igualmente, que todas as configurações possuem o mesmo número de dimensões p. Se, de facto, a i-ésima representação tem p_i dimensões e os p_i são diferentes podem tratar-se todas as configurações com dimensão igual ao maior dos p_i e colocando $(p-p_i)$, $p = \max p_i$, colunas de zeros na matriz $\mathbf{X}^{(i)}$.

Considere-se um ponto s em todas as configurações cujas coordenadas no i-ésimo gráfico são: $x_{s1}^{(i)}$, $x_{s2}^{(i)}$,..., $x_{sp}^{(i)}$. Se todas as representações têm em comum os p eixos ortogonais existe um conjunto de g pontos para cada entidade e, ao todo, n conjuntos deste tipo. O s-ésimo conjunto, correspondente ao s-ésimo ponto possui um centróide G_S de coordenadas ($\bar{x}_{s1}$, $\bar{x}_{s2}$,..., $\bar{x}_{sp}$) onde,

$$\bar{x}_{st} = \frac{1}{g} \sum_{i=1}^{g} x_{st}^{(i)} \quad , (t=1,\ldots,p) \quad . \tag{7.17}$$

Se as g representações se ajustam bem, cada conjunto de g pontos forma uma "nuvem" compacta. Se, pelo contrário, o ajustamento entre as g configurações é fraco, cada conjunto de g pontos tem um aspecto difuso. Um medida simples do ajustamento do s-ésimo conjunto de g pontos é a soma das distâncias quadradas entre cada ponto do conjunto e o centróide,

$$(7.18) \qquad \sum_{i=1}^{g} \sum_{t=1}^{p} \left(x_{st}^{(i)} - \bar{x}_{st} \right)^2 .$$

A extensão natural de M^2 para a análise generalizada de g representações é,

$$(7.19) \qquad \bar{M}^2 = \sum_{i=1}^{g} \sum_{s=1}^{n} \sum_{t=1}^{p} \left(x_{st}^{(i)} - \bar{x}_{st} \right) .$$

Esta medida é calculada após as configurações terem sofrido translações, rotações e mudanças de escala. Como foi referido atrás efectuar uma rotação a uma representação com coordenadas dadas pela matriz $\mathbf{X}^{(i)}$ é equivalente a pós-multiplicar $\mathbf{X}^{(i)}$ por uma matriz ortogonal $\mathbf{Q}^{(i)}$ e uma mudança de escala é expressa pela multiplicação por um escalar $c^{(i)}$. Translação para uma nova origem é conseguida através da adição de um mesmo vector linha $\mathbf{1}^{(i)}$ $(1xp)$ a cada linha de $\mathbf{X}^{(i)}$. Sendo $\mathbf{T}^{(i)}$ a matriz (nxp) que

contém os vectores $\mathbf{1}^{(i)}$ é possível expressar as operações através de $c^{(i)}\mathbf{X}^{(i)}\mathbf{Q}^{(i)}+\mathbf{T}^{(i)}$. A análise procrusteana generalizada pretende determinar $c^{(i)}$, $\mathbf{Q}^{(i)}$ e $\mathbf{T}^{(i)}$ de modo que $\bar{M}^2$ seja mínimo.

Gower(1975) recomenda sujeitar todas as matrizes $\mathbf{X}^{(i)}$ às operações impondo restrições para evitar soluções degeneradas tais como $c^{(i)}=0$. Ao efectuar as translações é conveniente tomar a origem dos eixos como centróide comum.

Neste método os passos que tratam da rotação e da mudança de escala não são independentes e a solução óptima em cada passo necessita do valor óptimo do outro. Assim a solução é obtida por um procedimento numérico iterativo descrito em pormenor por Gower(1975).

Para ilustrar a técnica procrusteana comparou-se a configuração resultante da aplicação de MDSC usando a distância euclideana com a representação obtida utilizando a mesma técnica de "*scaling*" mas em que a distância euclideana é expressa em percentagens e que constam, respectivamente, das figuras 4.2 e 4.7. Designou-se a primeira por **X** e a segunda por **Y**.

A análise procrusteana efectuada, composta por centragem e normalização de **X** e **Y** e rotação de **Y**, encontra-se na figura 7.2. A divergência entre as representações após a análise, é pequena,

facto comprovado pelo valor de M^2 = 0.2612.

Fez-se um ajustamento de escala que se representa na figura 7.3. O *c* óptimo é 0.8694 e com ele determina-se um M^2 de 0.2442, menor que o anterior.

Comparam-se, de seguida, os mapas das figuras 4.2 e 4.16. Este último resultou de uma análise de MDSO à matriz de opiniões de estudantes de economia. Efectuaram-se, todas as transformações com, nomeadamente, um factor de ajustamento de escala de 0.7106. A divergência entre as duas configurações está representada na figura 7.4 e é expressa por um valor de M^2 =0.4951.

FIGURA 7.2 – ANALISE PROCRUSTEANA entre os mapas das figuras 4.2 e 4.7

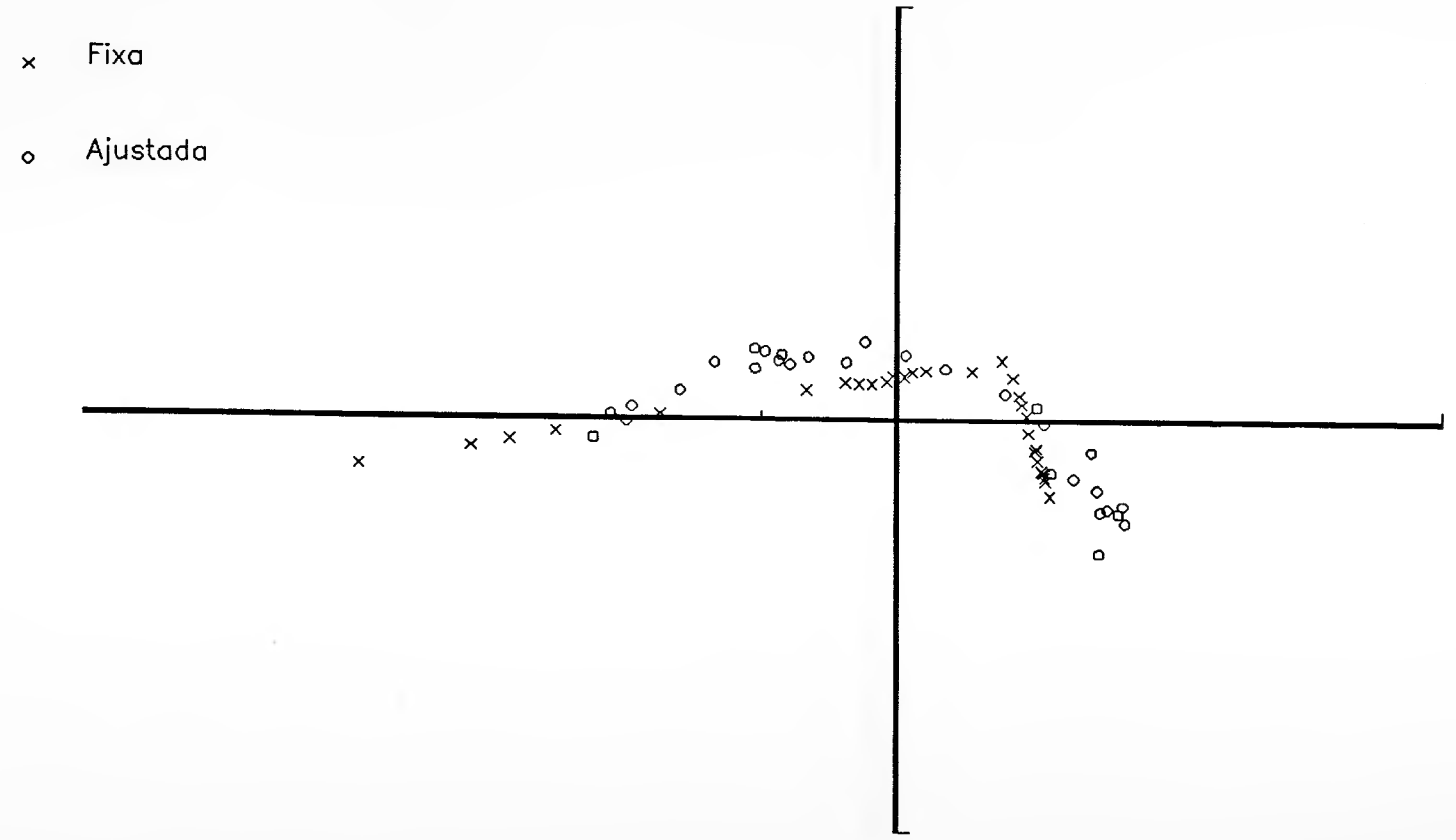


FIGURA 7.3 — ANALISE PROCRUSTEANA entre os mapas das figuras 4.2 e 4.7
— factor de escala = 0.8694 —

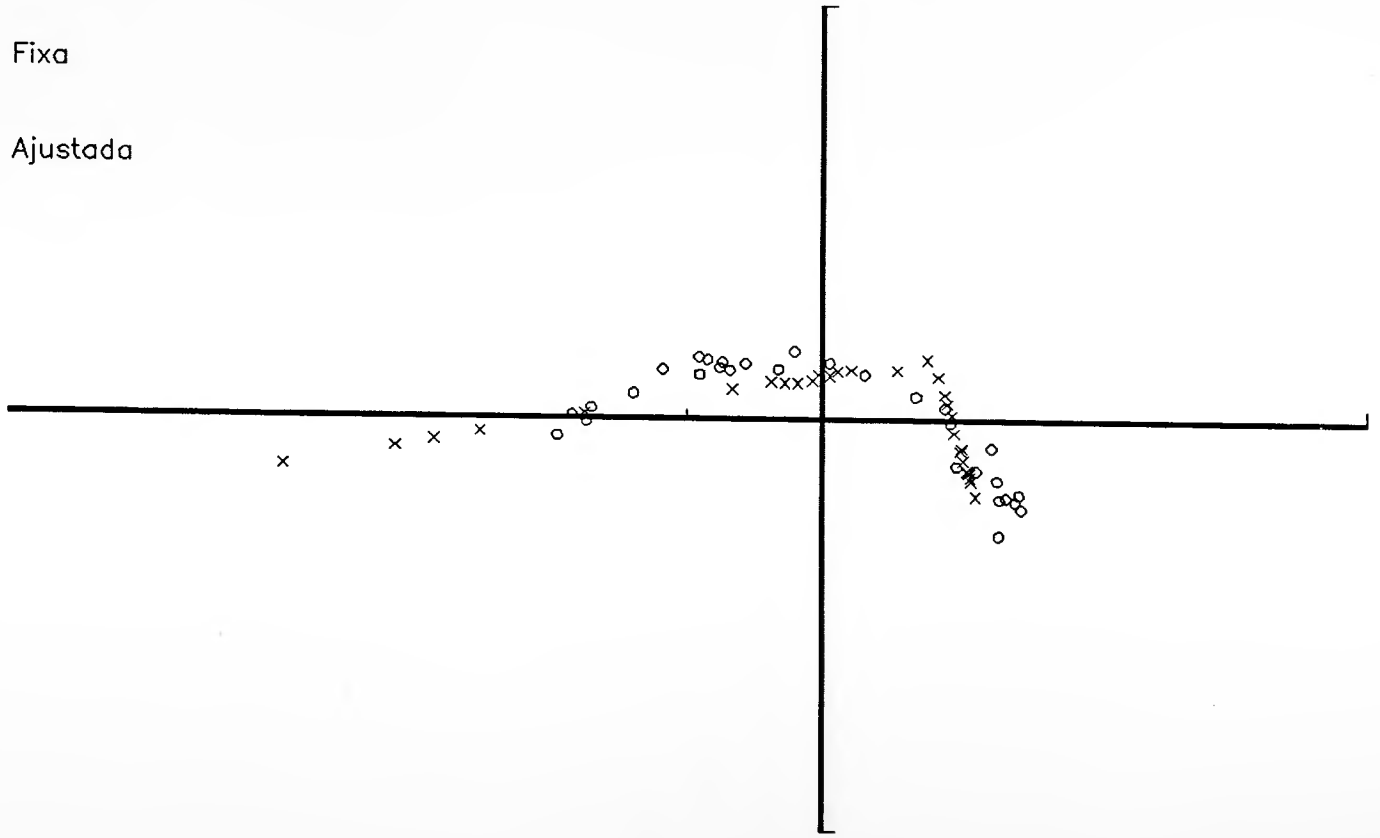
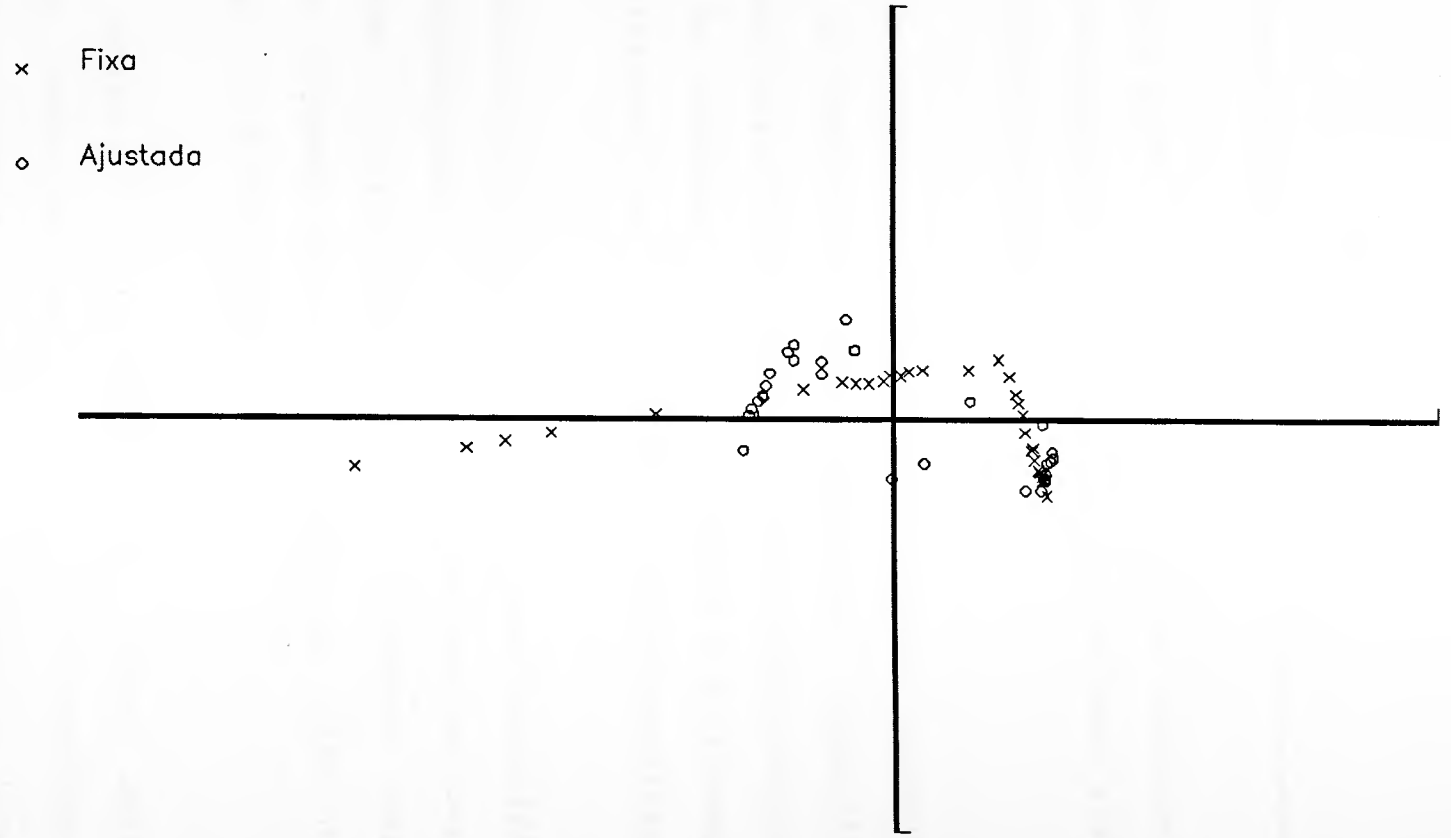


FIGURA 7.4 – ANALISE PROCRUSTEANA entre os mapas das figuras 4.2 e 4.16
– factor de escala = 0.7106 –



CAPÍTULO 8

CONCLUSÃO

Ao longo das páginas deste trabalho foram descritas técnicas que permitem a representação gráfica de dados multivariados visando uma exploração informal dos dados.

Os procedimentos de exploração informal destinam-se a desenvolver no investigador um 'sentimento' pelos dados que o capacite na antecipação de características da estrutura dos dados. Contrastam com os processos de inferência formal preocupados com afirmações, em geral probabilísticas, sob a regência de modelos específicos.

As técnicas de representação directa são uma solução simplificada para tratar a complexidade dos dados multivariados no caso de pequenas amostras. Confrontados com o pequeno esforço computacional que requerem dão, muitas vezes, uma visão útil e informativa da estrutura dos dados.

Os métodos de projecção e optimização que visam obter a representação dos dados num reduzido número de dimensões têm naturezas diferentes. MDSC e "*Biplot*" são procedimentos baseados

em valores e vectores próprios enquanto que MDSO e MDSD operam minimizando uma função particular através de algoritmos iterativos.

MDSC, MDSO e MDSD são técnicas poderosas aplicáveis directamente a dados que se apresentem sob a forma de matriz de proximidades. O MDSC possui a vantagem de não se limitar às dissemelhanças medidas com a distância euclideana, o MDSO emprega as propriedades ordinais das proximidades e o MDSD permite o uso de mais do que uma amostra para o fenómeno em estudo. O "*Biplot*" tem vantagens consideráveis quando se deseja representar simultaneamente as relações entre variáveis e sujeitos.

Nenhuma recomendação pode ser feita para eleger o melhor método pronto a usar em todas as situações. Em geral a escolha do modelo de representação gráfica é determinada por muitos factores, como sejam, o tipo de dados, o número de observações e a complexidade da estrutura esperada. Por vezes a utilização de mais do que um modelo pode ser ensaiada.

(3.2)

Para o ponto t_0 a diferença entre duas funções $f_{\mathbf{x}_l}$ e $f_{\mathbf{x}_k}$ é,

$$f_{\mathbf{x}_l}(t_0)-f_{\mathbf{x}_k}(t_0) \qquad (l,k=1,\ldots,i,\ldots,n).$$

O vector $\mathbf{x}_i$ tem coordenadas $\mathbf{x}_i = (x_{i1},\ldots, x_{ip})$ e t_0 assume qualquer valor entre $-\pi$ e π. No gráfico, a distância quadrática entre as duas funções $f_{\mathbf{x}_l}$ e $f_{\mathbf{x}_k}$ define-se,

$$\|f_{\mathbf{x}_l}-f_{\mathbf{x}_k}\|^2 = \int_{-\pi}^{\pi} [f_{\mathbf{x}_l}(t)-f_{\mathbf{x}_k}(t)]^2 dt =$$

$$= \int_{-\pi}^{\pi} [\ (\frac{1}{\sqrt{2}}x_{11} + x_{12}\text{sen } t + x_{13}\cos t + x_{14}\text{sen } 2t + x_{15}\cos 2t + \ldots)^2$$

$$- 2\ (\frac{1}{\sqrt{2}}x_{11} + x_{12}\text{sen } t + x_{13}\cos t + x_{14}\text{sen } 2t + x_{15}\cos 2t + \ldots)$$

$$(\frac{1}{\sqrt{2}}x_{k1} + x_{k2}\text{sen } t + x_{k3}\cos t + x_{k4}\text{sen } 2t + x_{k5}\cos 2t + \ldots)$$

$$+ (\frac{1}{\sqrt{2}}x_{k1} + x_{k2}\text{sen } t + x_{k3}\cos t + x_{k4}\text{sen } 2t + x_{k5}\cos 2t + \ldots)^2\ ]\ dt =$$

$$= \int_{-\pi}^{\pi} [\ \frac{1}{2}(x_{11}-x_{k1})^2 + \frac{2}{\sqrt{2}}(x_{11}-x_{k1})(x_{12}-x_{k2})\operatorname{sen} t + \frac{2}{\sqrt{2}}(x_{11}-x_{k1})$$

$$(x_{13}-x_{k3})\cos t + \frac{2}{\sqrt{2}}(x_{11}-x_{k1})(x_{14}-x_{k4})\operatorname{sen} 2t + \frac{2}{\sqrt{2}}(x_{11}-x_{k1})$$

$$(x_{15}-x_{k5})\cos 2t + \ldots + (x_{12}-x_{k2})^2\operatorname{sen}^2 t + 2(x_{12}-x_{k2})(x_{13}-x_{k3})$$

$$\operatorname{sen} t \cos t + 2(x_{12}-x_{k2})(x_{14}-x_{k4})\operatorname{sen} t \operatorname{sen} 2t + 2(x_{12}-x_{k2})$$

$$(x_{15}-x_{k5})\operatorname{sen} t \cos 2t + \ldots + (x_{13}-x_{k3})^2\cos^2 t + 2(x_{13}-x_{k3})$$

$$(x_{14}-x_{k4})\cos t \operatorname{sen} 2t + 2(x_{13}-x_{k3})(x_{15}-x_{k5})\cos t \cos 2t + \ldots +$$

$$(x_{14}-x_{k4})^2\operatorname{sen}^2 2t + 2(x_{14}-x_{k4})(x_{15}-x_{k5})\operatorname{sen} 2t \cos 2t + \ldots +$$

$$(x_{15}-x_{k5})^2\cos^2 2t + \ldots\ ]\ dt =$$

(As primitivas $F(t)$ de muitas das funções acima anulam-se quando calculadas para $F(\pi)-F(-\pi)$)

$$= \pi(x_{11}-x_{k1})^2 + \pi(x_{12}-x_{k2})^2 + \pi(x_{13}-x_{k3})^2 + \pi(x_{14}-x_{k4})^2 + \pi(x_{15}-x_{k5})^2 + \ldots =$$

$$= \pi\ [\sum_{j=1}^{p} (x_{1j}\ x_{kj})^2]\ .$$

(3.3)

$f_{\mathbf{x}}(t)$ é função linear nas p variáveis,

$$f_{\underline{\mathbf{x}}}(t) = \frac{1}{n\sqrt{2}} \sum_{i=1}^{n} x_{i1} + \frac{1}{n} \sum_{i=1}^{n} x_{i2}\operatorname{sen} t + \frac{1}{n} \sum_{i=1}^{n} x_{i3}\cos t$$

$$+ \frac{1}{n} \sum_{i=1}^{n} x_{i4} \operatorname{sen} 2t + \frac{1}{n} \sum_{i=1}^{n} x_{i5} \cos 2t + \ldots =$$

$$= \frac{1}{n} \sum_{i=1}^{n} f_{\mathbf{x}_i}(t) = \overline{f_{\mathbf{x}}(t)} \quad .$$

(3.4)

$$V[f_{\mathbf{x}}(t)] = V\left[\left(\frac{1}{\sqrt{2}}\mathbf{x}_1 + \mathbf{x}_2 \operatorname{sen} t + \mathbf{x}_3 \cos t + \mathbf{x}_4 \operatorname{sen} 2t + \mathbf{x}_5 \cos 2t + \ldots\right)\right] =$$

(Por hipótese, as p variáveis possuem variância comum σ^2 e não estão correlacionadas)

$$= \sigma^2 \left(\frac{1}{2} + \operatorname{sen}^2 t + \cos^2 t + \operatorname{sen}^2 2t + \cos^2 2t + \ldots\right) =$$

$$= \begin{cases} (1/2)\ \sigma^2\ p & (p \text{ ímpar}) \\ (1/2)\ \sigma^2\ [p - 1 + 2 \operatorname{sen}^2(pt/2) & (p \text{ par}). \end{cases}$$

(4.4)

$$d_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 =$$

$$= \sum_{k=1}^{p} [(x_{ik})^2 - 2\, x_{ik} x_{jk} + (x_{jk})^2] =$$

$$= \sum_{k=1}^{p} x_{ik} x_{ik} - 2 \sum_{k=1}^{p} x_{ik} x_{jk} + \sum_{k=1}^{p} x_{jk} x_{jk} =$$

$$= b_{ii} + b_{jj} - 2b_{ij}.$$

(4.5)

As coordenadas do centróide da configuração definem-se,

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik} \qquad (k=1,\ldots,p) \ .$$

Se o centróide se localiza na origem, as coordenadas $\bar{x}_k$ são 0, o que equivale a impor,

$$\sum_{i=1}^{n} x_{ik} = 0 \qquad (k=1,\ldots,p) \ .$$

(4.6)

Com a restrição (4.5) a soma dos elementos de **B** quer em linha quer em coluna é nula,

$$\sum_{j=1}^{n} b_{ij} = 0, \quad (i=1,\ldots,n) \ ,$$

e,

$$\sum_{i=1}^{n} b_{ij} = 0, \quad (j=1,\ldots,n).$$

Neste caso,

$$\sum_{i=1}^{n} d_{ij}^2 = \sum_{k=1}^{n} b_{kk} + n\, b_{jj} \quad ,$$

$$\sum_{j=1}^{n} d_{ij}^2 = \sum_{k=1}^{n} b_{kk} + n\, b_{ii} \quad ,$$

$$\sum_{i,j}^{n} d_{ij}^2 = 2n \sum_{k=1}^{n} b_{kk} \quad ,$$

$$\sum_{k=1}^{n} b_{kk} = \text{tr } \mathbf{B}.$$

Sendo,

$$d_{i.}^2 = \frac{1}{n} \sum_{j=1}^{n} d_{ij}^2 \quad ,$$

$$d_{.j}^2 = \frac{1}{n} \sum_{i=1}^{p} d_{ij}^2 \quad ,$$

$$d_{..}^2 = \frac{1}{n^2} \sum_{i=1}^{p} \sum_{j=1}^{p} d_{ij}^2 \quad ,$$

então,

$$b_{ij} = -\frac{1}{2} \left[d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2 \right] \quad .$$

(4.7)

A decomposição espectral de uma matriz simétrica ou decomposição em valores próprios é uma particularização de um resultado

poderoso designado decomposição em valores singulares. Qualquer matriz **X**, rectangular (nxp), pode ser decomposta no seguinte produto,

$$\mathbf{X} = \mathbf{U}\mathbf{T}\mathbf{V}^{T} ,$$

onde **U**, matriz (nxk), e **V**, matriz (pxk) com $k \leq \min(n,p)$, são ortogonais,

$$\mathbf{U}^{T}\mathbf{U} = \mathbf{U}\mathbf{U}^{T} = \mathbf{I},$$

e,

$$\mathbf{V}^{T}\mathbf{V} = \mathbf{V}\mathbf{V}^{T} = \mathbf{I},$$

devido à ortonormalidade dos vectores-coluna $\mathbf{u}_i$ (i=1,...,k) e $\mathbf{v}_j$ (j=1,...,k) que, respectivamente, constituem as matrizes **U** e **V**. Aos vectores $\mathbf{u}_i$ dá-se o nome de vectores singulares de **X** à esquerda e aos $\mathbf{v}_j$ dá-se o nome de vectores singulares de **X** à direita. A matriz **T** (kxk) é diagonal com elementos τ_s (s=1,...,k), reais, designados valores singulares de **X** que se encontram dispostos por ordem decrescente,

$$\tau_1 \geq \tau_2 \geq \ldots \geq \tau_k \geq 0.$$

Quando **X** é simétrica, por exemplo (nxn), verifica-se que **U** = **V** e a decomposição em valores singulares reduz-se à decomposição

espectral de $\mathbf{X}$,

$$\mathbf{X} = \mathbf{VLV}^T,$$

onde $\mathbf{V}$, matriz $(n \times n)$ é matriz ortogonal cujas colunas $\mathbf{v}_i$ $(i=1,\ldots,n)$ são os vectores próprios de $\mathbf{X}$, ortonormais e correspondentes aos valores próprios λ_i que constituem a matriz diagonal $\mathbf{L}$ $(n \times n)$e se encontram dispostos por ordem decrescente,

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n.$$

Ao contrário dos elementos τ_s que são positivos ou nulos, os λ_i podem assumir valores negativos.

Entre a decomposição de $\mathbf{X}$ em valores singulares e a decomposição espectral de $\mathbf{XX}^T$ e de $\mathbf{X}^T\mathbf{X}$ estabelecem-se relações interessantes,

$$\mathbf{XX}^T = \mathbf{UTV}^T\mathbf{VT}^T\mathbf{U}^T = \mathbf{UT}^2\mathbf{U}^T,$$

e,

$$\mathbf{X}^T\mathbf{X} = \mathbf{VT}^T\mathbf{U}^T\mathbf{UTV}^T = \mathbf{VT}^2\mathbf{V}^T.$$

Os vectores singulares de $\mathbf{X}$ à esquerda, $\mathbf{u}_i$, são os vectores próprios de $\mathbf{XX}^T$ e os vectores singulares de $\mathbf{X}$ à direita, $\mathbf{v}_j$, são os vectores próprios de $\mathbf{X}^T\mathbf{X}$. Os valores singulares de $\mathbf{X}$, τ_s $(s=1,\ldots,k)$ com $k \leq \min(n,p)$, constituem as raízes quadradas positivas dos valores próprios não negativos das duas matrizes

simétricas $\mathbf{X}\mathbf{X}^T$ e $\mathbf{X}^T\mathbf{X}$.

A decomposição em valores singulares possui diversas propriedades mas a mais importante, no contexto dos modelos geométricos, é a que reporta ao artigo de Eckart e Young(1936): uma matriz $\mathbf{X}$, de característica k, é aproximada, o melhor possível, por outra matriz $\mathbf{X}_{[t]}$ com a mesma dimensão mas de característica inferior, t, através da decomposição de $\mathbf{X}$ em valores singulares. A matriz $\mathbf{X}_{[t]}$ calcula-se efectuando o seguinte produto,

$$\mathbf{X}_{[t]} = \mathbf{U}_t\mathbf{T}_t\mathbf{V}_t^T \ ,$$

onde o índice t indica que as matrizes $\mathbf{U}_t$, $\mathbf{T}_t$ e $\mathbf{V}_t$ são formadas pelas primeiras colunas de $\mathbf{U}$, $\mathbf{T}$ e $\mathbf{V}$, respectivamente.

A melhor aproximação é entendida no sentido dos mínimos quadrados, o que significa que o somatório M das diferenças quadráticas entre elementos homólogos das matrizes $\mathbf{X}$ e $\mathbf{X}_{[t]}$ é mínimo,

$$M = \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - x_{[t]ij})^2 \ .$$

(4.8)

$\mathbf{L}$ é diagonal, com elementos todos positivos, então,

$$\mathbf{L} = \mathbf{L}^{1/2}\ \mathbf{L}^{1/2},$$

$$\mathbf{B} = \mathbf{X}\mathbf{X}^T = \mathbf{V}\mathbf{L}\mathbf{V}^T = \mathbf{V}\ \mathbf{L}^{1/2}\ \mathbf{L}^{1/2}\ \mathbf{V}^T,$$

pelo que,

$$\mathbf{X} = \mathbf{V}\ \mathbf{L}^{1/2} = [\ \sqrt{\lambda_1}\ \mathbf{v}_1 \quad \sqrt{\lambda_2}\ \mathbf{v}_2 \quad \ldots \quad \sqrt{\lambda_n}\ \mathbf{v}_n].$$

(4.9)

Sejam λ_j e $\mathbf{v}_j$ (j=1,...,p), respectivamente, os valores e vectores próprios de $\mathbf{X}^T\mathbf{X}$. Então,

$$(\mathbf{X}^T\mathbf{X})\ \mathbf{v}_j = \lambda_j \mathbf{v}_j\ .$$

Pré-multiplicando ambos os membros por $\mathbf{X}$,

$$(\mathbf{X}\mathbf{X}^T)\ \mathbf{X}\ \mathbf{v}_j = \lambda_j\ \mathbf{X}\ \mathbf{v}_j\ .$$

Esta equação mostra que os valores próprios de $\mathbf{X}\mathbf{X}^T$ são os mesmos de $\mathbf{X}^T\mathbf{X}$ e os vectores próprios de $\mathbf{X}\mathbf{X}^T$, designados por $\mathbf{u}_i$ (i=1,...,n), estão relacionados com os de $\mathbf{X}^T\mathbf{X}$ através de uma simples transformação linear: $\mathbf{u}_i$ são proporcionais a $\mathbf{v}_j$ por $\mathbf{X}$,

$$\mathbf{u}_i = \mathbf{X}\ \mathbf{v}_j.$$

(4.10)

A habitual medida da qualidade do ajustamento de um modelo através de mínimos quadrados é expressa pela proporção de variação em **X** que o modelo explica e traduz-se por,

$$G_1 = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - x_{[t]ij})^2}{\sum_{i=1}^{n}\sum_{j=1}^{p}x_{ij}^2} = \frac{Q - M}{Q} .$$

$$M = \sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - x_{[t]ij})^2 = tr\{(\mathbf{X}-\mathbf{X}_{[t]})(\mathbf{X}-\mathbf{X}_{[t]})^T\} =$$

$$= tr(\mathbf{X}\mathbf{X}^T) - 2\,tr(\mathbf{X}\mathbf{X}_{[t]}^T) + tr(\mathbf{X}_{[t]}\mathbf{X}_{[t]}^T)$$

$$= \sum_{i=1}^{n}\lambda_i - 2\sum_{i=1}^{t}\lambda_i + \sum_{i=1}^{t}\lambda_i = \sum_{i=t+1}^{n}\lambda_i .$$

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{p}x_{ij}^2 = tr(\mathbf{X}\mathbf{X}^T) = \sum_{i=1}^{n}\lambda_i .$$

$$G_1 = \frac{\sum_{i=1}^{n}\lambda_i - \sum_{i=t+1}^{n}\lambda_i}{\sum_{i=1}^{n}\lambda_i} = \frac{\sum_{i=1}^{t}\lambda_i}{\sum_{i=1}^{n}\lambda_i} .$$

(5.1)

Propriedades da característica de uma matriz,

i) Se $\mathbf{A}(nxp)$ a $c(\mathbf{A}) \leq \min\{n,p\}$

ii) $c(\mathbf{AB}) \leq \min\{c(\mathbf{A}), c(\mathbf{B})\}$

Se $\mathbf{Y}$ é (nxp) e a $c(\mathbf{Y})=k$ então $k \leq \min\{n,p\}$ ou seja $k \leq n$ e $k \leq p$;

$c(\mathbf{G}) \leq k$ e $c(\mathbf{H}^T)=c(\mathbf{H}) \leq k$;

mas $c(\mathbf{GH}^T)=k \leq \min\{c(\mathbf{G}), c(\mathbf{H})\}$ o que implica $c(\mathbf{G})=c(\mathbf{H})=k$.

(5.16)

$$G_S = 1 - \frac{\sum_{l=1}^{p}\sum_{j=1}^{p}(s_{lj} - \mathbf{h}_l\mathbf{h}_j^T)^2}{\sum_{l=1}^{p}\sum_{j=1}^{p} s_{lj}^2} = \frac{Q_S - M_S}{Q_S} .$$

$$M_S = \sum_{l=1}^{p}\sum_{j=1}^{p}(s_{lj} - \mathbf{h}_l\mathbf{h}_j^T)^2 = \operatorname{tr}\{(\mathbf{S}-\mathbf{HH}^T)(\mathbf{S}-\mathbf{HH}^T)^T\} =$$

$$= \operatorname{tr}(\mathbf{SS}^T) - 2\operatorname{tr}(\mathbf{SHH}^T) + \operatorname{tr}(\mathbf{HH}^T\mathbf{HH}^T)$$

$$= \sum_{j=1}^{p}[\tau_j^4/(n-1)^2] - 2\sum_{j=1}^{2}[\tau_j^4/(n-1)^2] + \sum_{j=1}^{2}[\tau_j^4/(n-1)^2]$$

$$= \sum_{j=3}^{p}[\tau_j^4/(n-1)^2] .$$

$$Q_S = \sum_{j=1}^{p}\sum_{j=1}^{p} s_{ij}^2 = \operatorname{tr}(\mathbf{SS}^T) = \sum_{j=1}^{p}[\tau_j^4/(n-1)^2] .$$

$$G_S = \frac{\sum_{j=1}^{2} [\tau_j^4 /(n-1)^2]}{\sum_{j=1}^{p} [\tau_j^4 /(n-1)^2]} = \frac{\sum_{j=1}^{2} \tau_j^4}{\sum_{j=1}^{p} \tau_j^4} \quad .$$

(7.3)

Soma-se e subtrai-se $\bar{x}_j$ e $\bar{y}_j$ em (7.1),

$$M^2 = \sum_{i=1}^{n} [\sum_{j=1}^{p} \{(x_{ij} - \bar{x}_j) - (y_{ij} - \bar{y}_j) + (\bar{x}_j - \bar{y}_j)\}^2] \quad ,$$

e tendo em atenção que ,

$$\sum_{i=1}^{n} (x_{ij} - \bar{x}_j) = \sum_{i=1}^{n} (y_{ij} - \bar{y}_j) = 0 \quad ,$$

então,

$$M^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} \{(x_{ij} - \bar{x}_j) - (y_{ij} - \bar{y}_j)\}^2 + n \sum_{j=1}^{p} (\bar{x}_j - \bar{y}_j)^2 .$$

(7.6)

$$M^2 = \sum_{i=1}^{n} [\sum_{j=1}^{p} (x_{ij} - y_{ij})^2 = \text{tr} \{ (\mathbf{X} - \mathbf{Y})(\mathbf{X} - \mathbf{Y})^T\} \quad ,$$

ou, desenvolvendo a expressão,

$$M^2 = \text{tr}\,(\mathbf{X}\mathbf{X}^T) + \text{tr}\,(\mathbf{Y}\mathbf{Y}^T) - 2\,\text{tr}\,(\mathbf{X}\mathbf{Y}^T) \quad ,$$

e substituindo $\mathbf{Y}$ por $\mathbf{Y}\mathbf{Q}$,

$$M^2 = \text{tr}\,(\mathbf{X}\mathbf{X}^T) + \text{tr}\,(\mathbf{Y}\mathbf{Y}^T) - 2\,\text{tr}\,(\mathbf{X}\mathbf{Q}^T\mathbf{Y}^T) \quad .$$

(7.7) e (7.8)

Lema (Sibson, 1978). Seja $\mathbf{A}$ uma matriz quadrada e $\mathbf{Q}$ uma matriz ortogonal da mesma dimensão. Então,

$$\text{tr}\,(\mathbf{Q}^T\mathbf{A}) \leq \text{tr}\,(\mathbf{A}^T\mathbf{A})^{1/2},$$

com igualdade se e só se $\mathbf{Q}$ satisfaz,

$$(\mathbf{Q}^T\mathbf{A}) = (\mathbf{A}^T\mathbf{A})^{1/2} \quad ,$$

onde $\mathbf{M}^{1/2}$ designa a raíz quadrada simétrica e não negativa da matriz simétrica e não negativa $\mathbf{M}$. Esta equação tem sempre uma solução $\mathbf{Q}$ ortogonal e se $\mathbf{A}$ é não singular a solução é única,

$$\mathbf{Q} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1/2} \quad .$$

Demonstração: A decomposição de uma matriz $\mathbf{A}$ em valores singulares é,

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$$

onde $\mathbf{U}$ e $\mathbf{V}$ são matrizes ortogonais e Σ é matriz quadrada diagonal contendo elementos não negativos na diagonal. Então,

$$\text{tr}\ (\mathbf{Q}^T\mathbf{A}) = \text{tr}\ (\mathbf{Q}^T\mathbf{U}\Sigma\mathbf{V}^T) = \text{tr}\ (\mathbf{V}^T\mathbf{Q}^T\mathbf{U}\Sigma) = \text{tr}\ (\mathbf{T}\Sigma),$$

onde $\mathbf{T} = \mathbf{V}^T\mathbf{Q}^T\mathbf{U}$ é matriz ortogonal, pelo que nenhum elemento de $\mathbf{T}$ é maior que 1 e,

$$\text{tr}\ (\mathbf{T}\Sigma) \leq \text{tr}\ (\Sigma)\ ,$$

com igualdade se e só,

$$\mathbf{V}^T\mathbf{Q}^T\mathbf{U}\Sigma = \Sigma.$$

A solução desta equação é,

$$\mathbf{Q}=\mathbf{U}\mathbf{V}^T,$$

que é a única solução se Σ é não singular. A equação pode também ser escrita,

$$\mathbf{Q}^T\mathbf{A} = \mathbf{V}\mathbf{U}^T\mathbf{U}\Sigma\mathbf{V}^T = \mathbf{V}\Sigma\mathbf{V}^T =$$
$$= (\mathbf{V}\Sigma^2\mathbf{V}^T)^{1/2} = (\mathbf{V}\Sigma\mathbf{U}^T\mathbf{U}\Sigma\mathbf{V}^T)^{1/2} =$$

$$= \{(U\Sigma V^T)^T(U\Sigma V^T)^{1/2}\}^{1/2} = (A^TA)^{1/2},$$

e a demonstração fica completa com ,

$$\text{tr } (\Sigma) = (VV^T\Sigma) = \text{tr } (V\Sigma V^T) = \text{tr } (A^TA)^{1/2}.$$

Obtém-se o seguinte teorema considerando $A = Y^TX$, na equação $(Q^TA) = (A^TA)^{1/2}$.

Teorema Se X e Y são configurações que foram centradas na origem M^2 é minimizado transformando Y em YQ onde Q é solução ortogonal de,

$$Q^TY^TX = (X^TYY^TX)^{1/2}.$$

Se Y^TX é não singular,

$$Q = Y^TX(X^TYY^TX)^{-1/2} ,$$

e o valor mínimo de M^2 é,

$$M^2 = \text{tr } (XX^T) + \text{tr } (YY^T) - 2 \text{ tr } (X^TYY^TX)^{1/2}.$$

BIBLIOGRAFIA

Anderson,E.(1960). A semigraphical method for the analysis of complex problems. *Technometrics*,**3**,387-91.

Anderson,T.W.(1958). *An Introduction to Multivariate Statistical Analysis*. Wiley,Nova York.

Andrews,D.F.(1972). Plots of high-dimensional data. *Biometrics*,**28**,125-36.

Andrews,D.F.(1980). Andrews function plots. In Johnson,N.L. e Kotz,S.(eds.) *Encyclopedia of Statistical Sciences*.Vol.1. Wiley,Nova York.

Andrews,D.F.,Gnanadesikan,R. e Warner,J.L.(1971). Transformations of multivariate data. *Biometrics*,**27**,825-40.

Bradu,D. e Gabriel,K.R.(1978). The biplot as a diagnostic tool for models of two-way tables. *Technometrics*,**20**,47-68.

Bruckner,L.A.(1978). On Chernoff faces. In Wang,P.C.C.(ed.) *Graphical Representation of Multivariate Data*. Academic Press,

Nova York.

Carroll,J.D. e Chang,J.J.(1970). Analysis of individual differences in multidimensional scaling via an *n*-way generalization of Eckart-Young decomposition. *Psychometrika*, **35**,283-319.

Chambers,J.M.,Cleveland,W.S.,Kleiner,B. e Tukey,P.A.(1983). *Graphical Methods for Data Analysis*. Wadsworth,Belmont.

Chatfield,C., e Collins,A.J.(1980). *Introduction to Multivariate Analysis*. University Press,Cambridge.

Chernoff,H.(1973). The use of faces to represent points in *k*-dimensional space graphically. *J. Am. Statist. Assoc.* **68**,361-68.

Chernoff,H. e Rizvi,M.H.(1975). Effect on classification error of random permutations of features in representing multivariate data by faces. *J. Amer. Statist. Assoc.*,**70**,548-54.

Corsten,L.A. e Gabriel, K.R.(1976). Graphical exploration in comparing variance matrices. *Biometrics*,**32**,851-63.

Cooley,W.W. e Lohnes,P.R.(1971). *Multivariate Data Analysis*. Wiley,Nova York.

Coxon,A.P.M.(1982). *The User's Guide to Multidimensional Scaling*. Heinemann Educational Books,Londres.

Du Toit,S., Steyn,G. e Stumpf,R.(1986). *Graphical Exploratory Data Analysis*. Springer-Verlag,Berlim.

Eckart,C. e Young,G.(1936).The approximation of one matrix by another of lower rank. *Psychometrika*,**1**,211-18.

Embrechts,P. e Herzberg,A.M.(1991). Variations of Andrews' plots. *International Statistical Revue*,**59**,175-94.

Everitt,B.S.(1978). *Graphical Tecniques in Multivariate Analysis*. Heinemann Educational Books,Londres.

Everitt,B.S.(1983). *Advanced Methods of data exploration and modelling*. Chapman and Hall,Londres.

Fienberg,S.E.(1979). Graphical methods in statistics. *The American Statistician*,Vol 33,**4**,165-78.

Flury,B. e Riedwyl,H.(1981). Graphical representation of multivariate data by means of asymmetrical faces. *J. Amer. Statist. Assoc.*,**76**,757-65.

Gabriel,K.R.(1971). The biplot-graphical display of matrices with

application to principal component analysis.*Biometrika*,**58**,453-67.

Gabriel,K.R.(1972). Analysis of meteorological data by means of canonical decomposition and biplots. *Journal of Applied Meteorology*,**11**,1071-7.

Gabriel,K.R.(1980). Biplot. In Johnson,N.L. e Kotz,S.(eds.) *Encyclopedia of Statistical Sciences*.Vol.1. Wiley,Nova York.

Gabriel,K.R.(1981). Biplot display of multivariate matrices for inspection of data and diagnosis. In Barnett,V.(ed.) *Interpreting multivariate data*. Wiley,Nova York.

Gabriel,K.R. e Zamir,S.(1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*,**21**,489-98.

Gnanadesikan,R. e Wilk,M.B.(1969). Data analytic methods in multivariate statistical analysis. In Krishnaiah,P.R.(ed) *Multivariate analysis II*. Academic Press,Nova York.

Gnanadesikan,R.(1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley,Nova York.

Goodchild,N.A. e Vijayan,K.(1974). Significance tests in plots of multidimensional data in two dimensions. *Biometrics*,**30**,209-10.

Gordon,A.D.(1981). *Classification*. Chapman and Hall,Londres.

Gower,J.C.(1966). Some distance properties of latent root and vector methods used in multivariate analysis.*Biometrika*,**53**,325-38.

Gower,J.C.(1971). A general coefficient of similarity and some of its properties. *Biometrics*,**27**,857-72.

Gower,J.C.(1975). Generalised Procrusts analysis. *Psychometrika*, **40**,33-51.

Gower,J.C.(1977). The analysis of asymmetry and orthogonality. In Barra,J. e outros(eds.) *Recent Developments in Statistics*. North-Holland,Amesterdão.

Gower,J.C. e Digby,P.G.N.(1981). Expressing complex relationships in two dimensions. In Barnett,V.(ed.) *Interpreting multivariate data*. Wiley,Londres.

Gower,J.C. e Ross,G.J.S.(1969). Minimum spanning trees and singles linkage cluster analysis. *Applied Statistics*.**18**,54-64.

Hartigan,J.A.(1985). Printer graphics for clustering. *J.Statist.Comput.Simul.* **4**,187-213.

Jacob,R.J.K.(1983). Investigating the space of Chernoff faces. In Rizvi,M.H.,Rustagi,J. e Siegmund,D.(eds) *Recent Advances in Statistics*. Academic Press,Nova York.

Kendall,Sir Maurice(1980).*Multivariate Analysis*.Griffin, Londres.

Kleiner,B. e Hartigan,J.A.(1981). Representing points in many dimensions by trees and castles (with discussion). *J.Am.Statist. Assoc.* **76**,260-76.

Kruskal,J.B.(1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*.**29**,1-27.

Kruskal,J.B.(1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*.**29**,115-29.

Kruskal,J.B. e Wish,M.(1978). *Multidimensinal Scaling*. Sage Publications,Londres.

Kulkarni,S.R. e Paranjape,S.R.(1984). Use of Andrews' function plot technique to construct control curves for multivariate process. *Commun.Statist.-Theor.Meth.*,**A13**,2511-33.

Krzanowski,W.J.(1990). *Principles of Multivariate Analysis:A User's Perspective*. Clarenden Press,Oxford.

Mardia,K.V.(1978). Some properties of classical multi-dimensional scaling. *Commun.Statist.-Theor.Meth.*,**A7**,1233-41.

Richardson,M.W.(1938). Multidimensional psychophysics. *Psychological Bulletin*,**35**,659-60.

Ross,G.J.S.(1972). Order invariante methods for data analysis (Discussion of Sibson,R.). *J.Roy.Statist.Soc.*,**B**,**34**,343-4.

Roy,S.N.(1957). *Some aspects of Multivariate Analysis*. Wiley,Nova York.

Santos,J.A.(1986). A lei de Wagner e a realidade das despesas públicas. *Estudos de Economia*,Vol VI,**2**,169-89.

Schönemann,P.H. e Carrol,R.M.(1970). Fitting one matrix to another under choice of a central dilation and rigid motion. *Psychometrika*,**35**,245-55.

Shepard,R.N.(1962a). The analysis of proximities: multidimensional scaling with an unknown distance function I. *Psychometrika*,**27**,125-40.

Shepard,R.N.(1962b). The analysis of proximities: multidimensional scaling with an unknown distance function II.

Psychometrika.**27**,219-46.

Sibson,R.(1978). Studies in the robustness of multidimensional scaling: procrustes statistics. *J.Roy.Statist.Soc.*,**40**,234-8.

Sibson,R.(1979). Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. *J.Roy. Statist.Soc.*,**41**,217-29.

Sibson,R.(1984). Present position and potencial development of multivariate analysis: some personal views. *J.R.Statist.Soc.A*, **147**,198-207.

Sneath,P.H.A. e Sokal,R.R(1973). *Numerical Taxonomy*. Freeman,São Francisco.

Spanier,J. e Oldham,K.B.(1987). *An Atlas of Functions*. Springer-Verlag,Berlim.

Takane,Y.,Young,F. e de Leeuw,J.(1977). Non-metric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features.*Psychometrika*,**42**,7-67.

Torgerson,W.S.(1958). *Theory and Methods of Scaling*. Wiley,Nova York.

Tukey,J.W. e Tuckey,P.A.(1981). Graphical display of data sets in three or more dimensions. In Barnett,V.(ed.) *Interpreting multivariate data*. Wiley,Nova York.

Wang,P.C.C.(1978) (ed.) *Graphical Representation of Multivariate Data*. Academic Press,Nova York.

Wang,P.C.C. e Lake,G.E.(1978). Graphical multivariate techniques in policy sciences. In Wang,P.C.C.(ed.) *Graphical Representation of Multivariate Data*. Academic Press,Nova York.

Wegman,E.J.,Carr,D.B. e Luo,Q.(1993). Visualizing multivariate data. In Rao,C.R.(ed.) *Multivariate Analysis: future directions*. North-Holland,Amesterdão.

Young,F.W.(1980). Multidimensional scaling. In Johnson,N.L. e Kotz,S.(eds.) *Encyclopedia of Statistical Sciences*.Vol.1. Wiley, Nova York.

Young,F.W. e Hamer,R.M.(1984). *Theory and Applications of Multidimensional Scaling*. Erlbaum Associates,Hillsdale.

Young,F.W. e Harris,D.F.(1990). Multidimensional scaling: Procedure ALSCAL. In *SPSS Base System User's Guide*. SPSS Inc.,Chicago.